REPORT RESUMES

ED 015 452 48 AL 000 802
A MATHEMATICAL ANALYSIS OF MULTI-LEVEL VERBAL LEARNING.
PSYCHOLOGY SERIES.
BY- BATCHELDER, WILLIAM H.
STANFORD UNIV., CALIF., INST.MATH.STUDIES SOC.SCI.
REPORT NUMBER TR-104 PUB DATE AUG 66
REPORT NUMBER BR-5-1088
CONTRACT OEC-6-14-009
EDRS PRICE MF-$0.75 HC-$6.72 166P.

DESCRIPTORS- *PSYCHOLOGY, EDUCATIONAL PSYCHOLOGY, PSYCHOMETRICS, *ASSOCIATIVE LEARNING, *VERBAL LEARNING, LEARNING PROCESSES, RECALL (PSYCHOLOGICAL), LEARNING THEORIES, TIME FACTORS (LEARNING), *MATHEMATICAL MODELS, RESEARCH METHODOLOGY, MARKOV MODELS, PAIRED ASSOCIATE LEARNING,

RESEARCH IS REPORTED HERE CONCERNING THE ASSUMPTION THAT ITEMS IN A LIST MUTUALLY AFFECT EACH OTHER IN VERBAL LIST-LEARNING. THE AUTHOR CONSIDERS BOTH THE MODE OF THE DATA ANALYSIS AND THE METHOD OF S-R (STIMULUS RESPONSE) PRESENTATION FOR A NUMBER OF RESTRICTED THEORETICAL ASSUMPTIONS INVOLVING ITEM INTERACTIONS IN S-R LIST-LEARNING EXPERIMENTS. IN SUCCEEDING CHAPTERS--(1) AN ANALYSIS IS MADE OF HOW BEST TO USE DATA IN LIST-LEARNING EXPERIMENTS SO AS TO BEAR ON A PSYCHOLOGICAL THEORY OR EVALUATE A MODEL--(2) A MODEL IS PRESENTED TO INDICATE HOW INFERENCES CAN BE MADE BY CONSIDERING A FIXED MODEL ON VARIOUS LEVELS OF DATA ANALYSIS AND IN VARIOUS EXPERIMENTAL SETTINGS ("ALL-OR-NONE MULTI-LEVEL MODEL")--(3) A MATHEMATICALLY RIGOROUS FRAMEWORK IS DEVELOPED FOR ANALYZING A LARGE CLASS OF MODELS WHICH EMBODY ITEM DEPENDENCIES--(4) THIS FRAMEWORK IS APPLIED TO SPECIFIC MODELS (INCLUDING THE "ALL-OR-NONE MODEL") AND RESULTS FOR VARIOUS PRESENTATION SCHEDULES ARE PRESENTED TO ILLUSTRATE THE FLEXIBILITY OF THE FRAMEWORK. ALSO ANALYZED IN TERMS OF THE FRAMEWORK ARE MARKOV MODELS DERIVED FOR A PARTICULAR CHOICE OF PRESENTATION SCHEDULE, AND RESTLE'S STRATEGY-SELECTION THEORY. THE FINAL CHAPTER REPORTS ON TWO LIST-LEARNING EXPERIMENTS CONDUCTED BY THE AUTHOR TO GENERATE DATA RELEVANT TO THESE IDEAS AND METHODS OF ANALYSES. (JD)

ED015452

BR-5-108
PA. 48

A MATHEMATICAL ANALYSIS OF MULTI-LEVEL VERBAL LEARNING

BY

WILLIAM H. BATCHELDER

TECHNICAL REPORT NO. 104
AUGUST 9, 1966

PSYCHOLOGY SERIES

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

AL 000 802

TECHNICAL REPORTS

PSYCHOLOGY SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

(Place of publication shown in parentheses; if published title is different from title of Technical Report, this is also shown in parentheses.)

1 D. Davidson, S. Siegel, and P. Suppes. Some experiments and related theory on the measurement of utility and subjective probability. August 15, 1955. (Experimental test of the basic model, Chapter 2 in Decision-making: An Experimental Approach. Stanford Univ. Press, 1957)

2 P. Suppes. Note on computing all optimal solutions of a dual linear programming problem. November 15, 1955.

3 D. Davidson and P. Suppes. Experimental measurement of utility by use of a linear programming model. April 2, 1956. (Experimental test of a linear programming model, Chapter 3 in Decision-making: An Experimental Approach. Stanford Univ. Press, 1957)

4 E. W. Adams and R. Fagot. A model of riskless choice. August 7, 1956. (Behavioral Science, 1959, 4, 1-10)

5 R. C. Atkinson. A comparison of three models for a Humphreys-type conditioning situation. November 20, 1956.

6 D. Scott and P. Suppes. Foundational aspects of theories of measurement. April 1, 1957. (J. Symbolic Logic, 1958, 23, 113-128)

7 M. Gerlach. Interval measurement of subjective magnitudes with subliminal differences. April 17, 1957.

8 R. C. Atkinson and P. Suppes. An analysis of two-person game situations in terms of statistical learning theory. April 25, 1957. (J. exp. Psychol., 1958, 55, 369-378)

9 R. C. Atkinson and P. Suppes. An analysis of a two-person interaction situation in terms of a Markov process. May 29, 1957. (In R. R. Bush and W. K. Estes (Eds.), Studies in Mathematical Learning Theory. Stanford Univ. Press, 1959. Pp. 65-75)

10 J. Popper and R. C. Atkinson. Discrimination learning in a verbal conditioning situation. July 15, 1957. (J. exp. Psychol., 1958, 56, 21-26)

11 P. Suppes and K. Walsh. A non-linear model for the experimental measurement of utility. August 21, 1956. (Behavioral Science, 1959, 4, 204-211)

12 E. Adams and S. Messick. An axiomatization of Thurstone's successive intervals and paired comparisons scaling models. September 9, 1957. (An axiomatic formulation and generalization of successive intervals scaling, Psychometrika, 1958, 23, 355-368)

13 R. Fagot. An ordered metric model of individual choice behavior. September 12, 1957. (A model for ordered metric scaling by comparison of intervals. Psychometrika, 1959, 24, 157-168)

14 H. Royden, P. Suppes, and K. Walsh. A model for the experimental measurement of the utility of gambling. September 25, 1957. (Behavioral Science, 1959, 4, 11-18)

15 P. Suppes. Two formal models for moral principles. November 1, 1957.

16 W. K. Estes and P. Suppes. Foundations of statistical learning theory, I. The linear model for simple learning. November 20, 1957. (Foundations of linear models. In R. R. Bush and W. K. Estes (Eds.), Studies in Mathematical Learning Theory. Stanford Univ. Press, 1959. Pp. 137-179)

17 D. Davidson and J. Marshak. Experimental tests of a stochastic decision theory. July 25, 1958. (In C. W. Churchman and P. Ratoosh (Eds.), Measurement: Definition and Theories. New York: Wiley, 1959. Pp. 233-269)

18 J. Lamperti and P. Suppes. Chains of infinite order and their application to learning theory. October 15, 1958. (Pacific Journal of Mathematics, 1959, 9, 739-754)

19 P. Suppes. A linear learning model for a continuum of responses. October 18, 1958. (In R. R. Bush and W. K. Estes (Eds.), Studies in Mathematical Learning Theory. Stanford Univ. Press, 1959. Pp. 400-414)

20 P. Suppes. Measurement, empirical meaningfulness and three-valued logic. December 29, 1958. (In C. West Churchman and P. Ratoosh (Eds.), Measurement: Definition and Theories. New York: Wiley, 1959. Pp. 129-143)

21 P. Suppes and R. C. Atkinson. Markov learning models for multiperson situations, I. The theory. February 20, 1959. (Chapter 1 in Markov Learning Models for Multiperson Interaction. Stanford Univ. Press, 1960)

22 J. Lamperti and P. Suppes. Some asymptotic properties of Luce's beta learning model. April 24, 1959. (Psychometrika, 1960, 25, 233-241)

23 P. Suppes. Behavioristic foundations of utility. July 27, 1959. (Econometrica, 1961, 29, 186-202)

24 P. Suppes and F. Krasne. Application of stimulus sampling theory to situations involving social pressure. September 10, 1959. (Psychol. Rev., 1961, 68, 46-59)

25 P. Suppes. Stimulus sampling theory for a continuum of responses. September 11, 1959. (In K. Arrow, S. Karlin, and P. Suppes (Eds.), Mathematical Methods in the Social Sciences. Stanford Univ. Press, 1960. Pp. 348-365)

26 W. K. Estes and P. Suppes. Foundations of statistical learning theory, II. The stimulus sampling model. October 22, 1959.

27 P. Suppes and R. C. Atkinson. Markov learning models for multiperson situations, II. Methods of analysis. December 28, 1959. (Chapter 2 in Markov Learning Models for Multiperson Interactions. Stanford Univ. Press, 1960)

28 R. C. Atkinson. The use of models in experimental psychology. May 24, 1960. (Synthese, 1960, 12, 162-171)

29 R. C. Atkinson. A generalization of stimulus sampling theory. June 14, 1960. (Psychometrika, 1961, 26, 281-290)

30 P. Suppes and J. M. Carlsmith. Experimental analysis of a duopoly situation from the standpoint of mathematical learning theory. June 17, 1960. (International Economic Review, 1962, 3, 1-19)

31 G. Bower. Properties of the one-element model as applied to paired-associate learning. June 29, 1960. (Application of a model to paired-associate learning, Psychometrika, 1961, 26, 255-280)

32 J. H. Blau. The combining of classes condition in learning theory. August 23, 1960. (See Transformation of probabilities, Proceedings of the Amer. Math. Soc., 1961, 12, 511-518)

33 P. Suppes. A comparison of the meaning and uses of models in mathematics and the empirical sciences. August 25, 1960. (Synthese, 1960, 12, 287-301)

34 P. Suppes and J. Zinnes. Stochastic learning theories for a response continuum with non-determinate reinforcement. October 25, 1960. (Psychometrika, 1961, 26, 373-390)

35 P. Suppes and R. Ginsberg. Application of a stimulus sampling model to children's concept formation of binary numbers, with and without an overt correction response. December 14, 1960. (Application of a stimulus sampling model to children's concept formation with and without an overt correction response, Journal exp. Psychol, 1962, 63, 330-336)

(Continued on inside back cover)

A MATHEMATICAL ANALYSIS OF MULTI-LEVEL VERBAL LEARNING

by

William H. Batchelder

TECHNICAL REPORT NO. 104

August 9, 1966

PSYCHOLOGY SERIES

Reproduction in Whole or in Part is Permitted for
any Purpose of the United States Government

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

ACKNOWLEDGMENTS

Acknowledgment is gratefully made to Professor Edward J. Crothers for his frequent advice and encouragement in the course of this dissertation, for which he served as chairman. Professor James G. Greeno also offered valuable advice throughout this investigation. Thanks are due also to Professors William K. Estes and Richard C. Atkinson, who were the other members of the dissertation committee.

Several other people have given valuable aid in the project. The writer has profited from many discussions with David E. Rumelhart, a research colleague. Also, Professor Patrick Suppes has given support and encouragement throughout the project. Thanks are due David Wessel for preparing the figures, and Karen Oxendine and Ruth Korb for typing the final manuscript. Finally, many thanks to my wife Margie for providing encouragement and for typing portions of the rough draft.

Work on this dissertation was begun during the author's tenure as a United States Public Health Terminal Year Predoctoral Fellow. The research was supported by Contract OE 6-14-009 from the United States Office of Education.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS iii

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER

1 INTRODUCTION 1

2 DATA ANALYSIS ON VARIOUS LEVELS 11

Comparison of One-Element P-Level and R-Level Models . . . 13

Model-Free Analyses of P- and R-Level Learning 21

3 THE ALL-OR-NONE MULTI-LEVEL MODEL 29

Axioms for the Model 29

Theorems and Derivations 32

Derivations for the Anticipation Procedure 36

4 GENERAL FRAMEWORK FOR MULTI-LEVEL MODELS 43

Reformulation of the All-or-None Multi-Level Model 45

The Framework 53

A. History of Major Ideas 53

B. Definition of Model in Framework 55

C. Definitions and Theorems for Presentation Schedules . . 61

D. Reduction of the Cardinality of Models 71

5 APPLICATIONS OF THE FRAMEWORK 91

The Mixed Model 91

The All-or-None Multi-Level Model (M = 2) 98

Another Version of the All-or-None Multi-Level Model . . . 106

Strategy Selection Theory 108

TABLE OF CONTENTS (cont.)

CHAPTER Page

6 EXPERIMENTS AND CONCLUSIONS 118

Experiments 118

Experiment I 120

Method 120

Results & Discussion 123

Experiment II 137

Method 137

Results & Discussion 138

APPENDIX I 149

REFERENCES 152

LIST OF TABLES

Table Page

2.1 Comparison of (P,P) and (P,R) 18

2.2 Comparison of (R,R) and (R,P) 20

3.1 Conditions under Which the All-or-None Multi-Level Model Reduces to the Four Analyses of the Preceding Chapter . . . 37

3.2 P- and R-level Analyses of the All-or-None Multi-Level Model 39

6.1 Mean Total Errors, E(T), and Mean Trial Number of Last Error, E(L), for C_3 and C_0 (List I) 124

6.2 Mean Total Errors, E(T), and Mean Trial Number of Last Error, E(L), for C_3, C_2^+, C_2^-, and C_0 (Lists II and III) 132

6.3 Mean Total Errors, E(T), and Mean Trial Number of Last Error, E(L), for C_3^+, C_3^- and Lists II, III 141

6.4 $Pr(x_n = 1 | L > n)$ for C_3^+, C_3^- and Lists II, III 145

6.5 $Pr(x_{n+1} = 1 | x_n = 1)$ for C_3^+, C_3^- and Lists II, III 145

LIST OF FIGURES

Figure Page

2.1 R-level Learning Curve for the List Depicted on p. 25 $(M = 4)$ 27

3.1 R-level Learning Curve for All-or-None Multi-level Model $(g = 1/5,\ r = 0.1,\ p = 0.3,\ M = 3)$ 41

5.1 Probability of Correct to S_2 for Mixed Model, $\pi = .95,\ c = .50$ 97

6.1 P-level Learning Curves for C_0 and C_3, List I 125

6.2 Total Error Distributions, T, and Trial Number of Last Error Distributions, L, for C_3, C_0, List I 126

6.3 & 6.4 Pr(error on n prior to last error) and Pr(error on n+1|error on n), respectively, for C_3 and C_0 (List I) 127

6.5 P-level Learning Curves for C_3 & C_0 and C_2^+ & C_2^- for Lists II, III 130

6.6 Total Error Distributions for C_3, C_2^+, C_2^-, C_0 for Lists II, III 131

6.7 Section of the R-level Learning Curve for C_3 for Lists II, III 135

6.8 P-level Learning Curve for C_0 and C_3 for List I (Exp. II) 139

6.9 P-level Learning Curves for C_3^- and C_3^+, Lists II, III (Exp. II) 140

6.10 Distributions of Total Errors, T, for C_3^+ and C_3^-, Lists II, III (Exp. II) 142

6.11 Distributions of the Trial Number of Last Error, L, for C_3^+ and C_3^-, Lists II, III (Exp. II) 143

6.12 R-level Learning Curves for the First Three Cycles on C_3^+ and C_3^- for Lists II, III 147

CHAPTER I

INTRODUCTION

This paper considers some of the implications of the assumption that items in a list mutually affect each other in the course of verbal list-learning. By a mutual affect or item interaction (or item dependency) is meant that performance on a particular S-R pair in a list depends in some way on the number and order of presentations of other S-R pairs in the list. It is hardly necessary to document the fact that items do interact in this sense; other things being equal, more errors are made to a particular S-R pair the larger the number of other S-R pairs in the list. Of course, these item interactions may be of a mild and uniform sort, such as might be produced by the subject's spreading his effort over M items, rather than just one; or the interactions might be more extreme and non-uniform, such as those postulated by a concept-identification model (cf. Restle, 1961). We open the analysis by drawing some conclusions from a brief review of the history of mathematical learning models for verbal list-learning.

Probably the simplest mathematical model for verbal list-learning is the one-element pattern model (Estes, 1959). This model was first analyzed in depth and applied to paired-associate learning data by Bower (1960, 1961). Since its introduction, the one-element model has received a number of diverse interpretations; among these are the following: (1) a stimulus pattern interpretation (Estes, 1959), (2) an all-or-none strategy-selection (hypothesis) interpretation (Restle, 1961, 1964), (3) a memory interpretation (Atkinson, Bower, and Crothers, 1965, pp. 87-88), and

(4) a response-elimination interpretation (Millward, 1964). Although the model has been successful in accounting for some list-learning data, a number of deficiencies in the model have been pointed out. Some of these are as follows: (1) the model fails to account for individual differences and unequal item difficulty (Postman, 1963), (2) learning may involve more than one stage (Restle, 1964), and (3) improvement in performance may take place prior to the last error (Suppes and Ginsberg, 1963).

Despite the ups and downs of the one-element model and its many modifications and extensions, the basic research strategy depicted in Bower (1961) has had a great influence on later invention and application of models to paired-associate data. This strategy has been first to state a (new) mathematical model for paired-associate learning (usually some finite-state Markov model), derive a battery of statistics from this model, estimate parameters in the model, and then attempt to account for summary statistics of the pool of subject-item error-success sequences[1] obtained in a list-learning experiment (usually run by the anticipation procedure) designed to validate the model. Among the many research papers exhibiting this four-step strategy are Atkinson and Crothers (1964), Bower (1961), Bower and Theois (1964), Calfee and Atkinson (1965), Millward (1964), Norman (1964), Polson, Restle, and Polson (1965), and Restle (1964).

A few of the models presented in these references have psychological rationals which assume that the learning of a particular S-R pair procedes independently of the states and responses of other items in the list

---

1/ Suppose a subject learns a list of M items by the anticipation procedure. Then that subject contributes M error-success subsequences, one for each item, to the pool of subject-item error-success sequences.

(e.g., the one-element model (Bower, 1961) and the random-trial increments model (Norman, 1964)). For these models it seems reasonable to study separate error-success sequences for each subject-item, since the model assumes that each of these sequences represents an independent sample path from the stochastic process represented in the model. However, for a number of other models, this independence of subject-item error-success sequences is placed in immediate question by the psychological theory postulated to underlie the model which is applied to these sequences. For example, the trial-differential-forgetting (T.D.F.) model suggested by Atkinson and Crothers (1964) and developed in Calfee and Atkinson (1965) postulates that the more intervening unlearned items between two successive presentations of a particular S-R pair, the greater the chance that the pair passes out of the short-term memory state and is thus forgotten. This assumption very definitely implies that, for a particular subject, error-success protocols for each item are not independent. Also, the strategy-selection theory of Restle (1964) implies that confusable items produce very non-independent error-success protocols, i.e., if $S\text{-}R_1$ and $S'\text{-}R_2$ are two such S-R pairs, the error-success process on each should be related, since subjects may confuse S and S'.

At best, an application of these models to a pool of error-success protocols which lack a stimulus tag or a subject tag represents an approximation to the true state of affairs. When applying the T.D.F. model to data (Calfee and Atkinson, 1965), it is assumed that the average number of unlearned items, F_n, intervening between the n^{th} and $n+1^{st}$ presentations of a given item applies to all items in a list. Under this approximation, the theory takes the form of a finite-state inhomogeneous

Markov chain. This chain is designed to account for the error-success protocols for each subject-item in the experiment. The approximation that Restle uses to account for subject-item protocols is discussed in detail in Chapter 5, pp. 108-117 of this paper. Basically, he neglects the interrelationships between a pair of confusable items in his applications.

The major psychological ideas in these latter two theories are as follows: (1) the T.D.F. model is based on the idea that unlearned items, when they are presented, cause items in short-term memory to be bumped into a forgotten state; and (2) the strategy-selection theory is based on the idea that stimulus confusion $(S\text{-}R_1, S'\text{-}R_2)$ is overcome in an all-or-none manner. In both cases we have seen that in order to apply the theory to a pool of subject-item error-success sequences in an anticipation procedure experiment, the major new variable in the theory is represented as an "average" quantity. However, by their nature, both the memory assumption and the confusion assumption imply highly differential effects on response probability to a particular S-R pair as a function of the number and order of other preceding S-R pairs. The implications of these two assumptions can be powerfully tested by either designing an experiment where S-R presentations are highly controlled or by utilizing statistics in the data that relate performance on separate items (or both possibilities together). Experiments and analyses of this nature have been performed on the memory assumption (Bjork, unpublished doctoral dissertation; Greeno, 1966; and Atkinson and Shiffrin, 1965) and on the stimulus confusion assumption (Restle, 1964, pp. 145-160; Ruskin, unpublished doctoral dissertation; and Sheppard, Hoveland, and

Jenkins, 1961). Finally, it should be mentioned that although the T.D.F. model and strategy-selection theory were singled out as being convenient examples of approaches to item dependencies, other models have also attempted to handle this problem.

This paper considers both the mode of data analysis and the method of S-R presentation for a number of restricted theoretical assumptions involving item interactions in S-R list-learning experiments. Chapter 2 considers the problem of level of data analysis, i.e., the problem of how to use data in a list-learning experiment to bear on a psychological theory or to evaluate a model. By this concept is meant the following: each subject in a list-learning experiment can be conceptualized as emitting a single finite data sequence. A particular member, x_n, of this sequence consists of the stimulus presented to the subject on the n^{th} trial, S_n, and his response to that stimulus, A_n. Thus, for a given subject i, the data are of the form

$$\begin{aligned} x^i &= x_1^i x_2^i \dots x_n^i \dots x_N^i \\ &= S_1^i A_1^i \; S_2^i A_2^i \dots S_n^i A_n^i \dots S_N^i A_N^i , \end{aligned} \tag{1.1}$$

where N S-R presentations are given to subject i in the experiment. In order to analyze data in a list-learning experiment, researchers transform this primary datum in ways to extract what they regard as its informative aspects. For example, for a subject-item error-success analysis, the primary datum is separated into subsequences, one for each item, and then the SA terms in these shorter sequences are transcribed as errors or successes. The particular way in which the primary datum is reduced represents the level of analysis.

More specifically, by level of data analysis is meant the collection of stimulus classes that define the error-success subsequences used in an analysis. With each class there is associated a single error-success sequence consisting of the chronological record of responses to members of the class. The subject-item analysis or paired-associate level (P-level) analysis consists of singleton stimulus classes, i.e., one for each item. On the other hand, a concept-level analysis (Atkinson, Bower, and Crothers, 1965, pp. 30-31) groups all stimuli in a list to define a single stimulus class giving rise to one error-success protocol for each subject. The units of a given level are the particular stimulus classes, e.g., for a P-level analysis, the units are the individual items. Another level of analysis discussed in Chapter 2 is as follows. Suppose a list of $J \cdot M$ S-R pairs is composed of J classes of M S-R pairs, where the items in any class of M items are interrelated and paired with the same response. The rule level (R-level) of analysis is defined to be the analysis where each group of M stimuli forms a stimulus class which defines a single error-success sequence for the class. Thus each subject would donate J error-success sequences for an R-level analysis. The units for this analysis would be the J classes of stimuli. Chapter 2 discusses methods of drawing inferences from a model (or psychological theory) by investigating alternative levels of analysis on the same set of data.

Chapter 3 extends Chapter 2 in the following sense: while much of Chapter 2 concerns the one-element model, Chapter 3 presents a model which is analogous but which allows subjects to learn either a particular S-R pair or a collection of related S-R pairs on a particular trial.

This model is called the all-or-none multi-level model, since it assumes that learning can take place at two levels simultaneously. These two levels are the P-level, corresponding to a P-level data analysis, and the R-level, corresponding to an R-level data analysis. Alternatives to the paired-associate anticipation procedure (i.e., the procedure whereby random permutations of the entire list are presented sequentially) are introduced, and some of the implications of the all-or-none multi-level model for these experimental procedures are presented. The models discussed in Chapters 2 and 3 are not designed to represent a theory of paired-associate learning but to indicate how inferences can be made by considering a fixed model on various levels of data analysis and in various experimental settings.

Chapter 4 establishes a mathematically rigorous basis for analyzing a large class of models which embody item dependencies. In this formulation, each item is allowed to have a different effect on the states of all the items in the list when it is presented for an anticipation trial; and, further, the state of each unpresented item in the list can effect the response probabilities and transition probabilities of the presented item. Among the motivations for developing this general mathematical framework are the following:

(1) The analysis of the all-or-none multi-level model in Chapter 3 is limited, owing to the difficulty in deriving properties of the model from the axiomatization presented in that chapter. The formulation of the all-or-none multi-level model in Chapter 3 is along the lines that models are conventionally axiomatized in the literature (cf. Atkinson, Bower, and Crothers, 1965, p. 85 and p. 353);

namely, a particular item is singled out and the various things which the model postulates can happen to that item are presented. At the outset of Chapter 4, the argument is made that when a model postulates item interactions, it might be more profitably analyzed in the context of a set of axioms that describe the things that can happen to the whole list of S-R pairs upon a presentation of a particular item. The chapter then develops this analysis and demonstrates that it helps overcome analytical difficulties that were inherent in the single-item axiomatization.

(2) An increasing number of mathematical models for list learning are embodying processes which involve item dependencies. Therefore, such models might profit from an analysis in terms of a framework designed to handle these dependencies. The argument for this case is presented in more detail in Chapter 4, pp. 49-53.

(3) Many experimenters have argued that most current list-learning experiments involve processes which concern interrelationships between items during the course of learning. Investigators have discovered a variety of psychological processes which operate, in varying degree, in such experiments. Most notable are the following processes: (i) memory and its organization (cf. Peterson, 1963; Melton, 1963), (ii) coding processes (cf. Symposium on coding and conceptual processes in verbal learning, articles by Battig, Cohen, Cofer, Tulving, Kendler, Shepard, 1966), and (iii) in second-language learning, dependencies arising either because of transfer from English or because of linguistic dependencies that are built into the second language (Crothers and Suppes, in press). The for-

mulation in Chapter 4 is designed to handle models which postulate processes like these and others which are similar.

(4) Traditionally, mathematical learning models have not been stated on levels that are general enough to constitute theories of paired-associate learning. By this is meant that many of the learning models are designed to predict performance only for a particular experimental procedure and level of data analysis. A model formalized in the framework of Chapter 4 can, in principle, predict performance for any mode of S-R presentation chosen for experimentation. The stochastic process which predicts performance for a particular presentation schedule comes as a logically tight derivation from the theory and does not represent the theory itself. Examples of how a stochastic model is derived from a general learning model axiomatized in the framework of Chapter 4 are presented in Chapter 5, pp. 103,105,115.

(5) Another contributer to the motivation for including Chapter 4 is the bias that progress in mathematical learning theory need not always be made by proposing a new theory of verbal learning (this is not attempted in the paper) but by the bringing of formal tools to the task of constructing new methods for drawing inferences from data (for example, the correlational analyses developed in Chapter 2, pp. 22 - 24) as well as constructing a formal framework for drawing conclusions from a theory once it is stated.

For these reasons, it is felt that Chapter 4 represents a definite contribution to mathematical learning theory, over and above the more specific developments in the other chapters. Nonetheless, the contribution does not represent a final solution to the problems we have raised.

Of course, these problems and observations which motivated Chapter 4 had been previously recognized by other investigators, and they are pooled to warn the reader of the particular bent that the paper (especially Chapter 4) will take.

Chapter 5 illustrates how the framework of Chapter 4 can be applied to specific models. An analysis of the mixed model paralleling that of Atkinson and Estes (1963) is presented in terms of the framework. Results for various presentation schedules are presented to illustrate the flexibility of the framework. Next the all-or-none multi-level model receives an additional analysis (to that given in Chapter 3) in terms of the framework. The additional feature of this analysis is that the process of deriving Markov models for a particular choice of presentation schedule is illustrated (Chapter 5, pp. 103,105). Finally, Restle's strategy-selection theory is developed in terms of the framework, and several problems with its earlier axiomatizations are met squarely by this analysis.

In Chapter 6, several experiments that the writer has conducted are briefly discussed. In addition, possible directions for further experimentation and analysis of multi-level processes in list-learning are indicated.

CHAPTER 2

DATA ANALYSIS ON VARIOUS LEVELS

In this chapter an analysis of the problem of levels of learning is initiated in a somewhat restrictive situation. Suppose one has a list of S-R pairs to be presented to subjects by the anticipation procedure. Assume the list is structured so that groups of stimuli paired to the same response have inter-relationships, e.g., all stimuli paired to a certain response start with the letter A, or all stimuli in a certain response class are names of animals.

For ease of presentation it will be assumed, for the moment, that learning is either on the single item level (P-level), on the rule level (R-level), or on both. Let us illustrate this with the following list:

Stimulus	Response
LEBESGUE	1
RIEMANN	1
STIELTJES	1
FISHER	2
BENKO	2
RESHEVSKY	2
STICKLES	3
PARKS	3
CASEY	3

Depending on instructions and whether or not the subject is familiar with mathematicians responsible for a method of integration, contemporary American chess players, or offensive ends for the San Francisco Forty-Niners, respectively, the subject might learn single S-R pairs or groups of S-R pairs. The assumption made in this chapter is that the unit of learning is the single item (P-level), or the set of 3 items related by the rule (R-level).

The concern of this chapter is with the properties of performance measures, i.e., the predictions a learning model might make for various ways of viewing the performance data. Considering only errors and successes, the primary datum from a subject in an anticipation procedure experiment is a long string of stimulus-result (error or success) pairs. The two modes of data analysis corresponding to the theoretical notions of P and R-level learning are as follows: 1. For a P-level analysis we abstract and pool all subsequences from the primary datum corresponding to each stimulus in the list; and 2. For an R-level analysis we abstract and pool all subsequences corresponding to a particular response. The example list has 9 P-level subsequences and 3 R-level subsequences for each subject.

In the literature, models are usually developed with a particular level of data analysis in mind (e.g., Bower, 1961, Restle, 1961). Even so, a model can be viewed as a stochastic process which generates sequences of 1s and 0s (errors and successes). If one wishes to apply a model, viewed in this way, to his data, he must choose a level (or levels) on which to apply it (e.g., Suppes, Crothers, Weir, and Trager, 1962).

Any nontrivial learning model[2] predicts that data will look different when analyzed on different levels. As a proof consider a primary datum as a string of 1s and 0s. Depending on which subsequences are abstracted for analysis, different results on such statistics as, for example, the proportion of 1s in the fifth place (i.e., Pr(error on "trial" 5)) are likely.

---

[2] Of course, for a trivial model producing strings of all zeros, each subsequence would also consist of all zeros and provide a single counter example.

It is a logical possibility that two learning models could agree in predictions on one level of analysis but disagree on another. To see this possibility consider the primary datum of strings of 1s and 0s, two models might agree on the probability distributions over subsequences but non-independence considerations might cause them to disagree on distributions over the primary datum level. For example it is possible that a simple model could fit P-level data and yet fail to account for an R-level analysis of the same data. Finally, it is possible for a choice of models to be correct but a choice of level of analysis to be wrong. Such a possibility must have occurred to Suppes, et. al. (1962) who actually used the same model on several levels of data analysis.

Comparison of One-element P-level and R-level Models

In this section we shall investigate the implications of the one-element model holding on either the R-level or the P-level. To illustrate some of the points above, both the R-level and P-level analysis of data generated by the P-level and R-level one-element model will be presented. In the next chapter a model allowing both types of learning will be presented.

The one-element model to be used in this analysis takes the following form (Estes, 1959, Bower, 1961). The unit to be learned starts in an unlearned state U. On the presentation of a unit in state U, the correct response is made with probability g and an error with probability 1-g. After response, the unit shifts to a learned state L with probability c and remains in U with probability 1-c. Units in L are always responded to correctly; and, once in L, a unit remains there. These assumptions are conveniently summarized by the transition matrix for the implied two-state Markov chain:

$$\begin{array}{ccc} & \text{state on trial } n+1 & \Pr(\text{correct}\mid\text{row state}) \\ & \begin{array}{cc} L & \quad U \end{array} & \\ \text{state on trial } n \;\; \begin{array}{c} L \\ U \end{array} & \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix} & \begin{bmatrix} 1 \\ g \end{bmatrix}. \end{array}$$

If the unit is a single item, we shall refer to the model as the one-element P-level model. If the unit is a group of items paired with the same response, we shall refer to the model as the one-element R-level model. Logically there are four possibilities for jointly considering the level of data analysis and the type of one-element model. These are (P,P), (P,R), (R,P), and (R,R), where the first letter refers to the level that data statistics are examined and the second identifies the model.

The (P,P) and (R,R) analyses are analogous to the usual paired-associate analysis of the one-element model (Bower, 1961) and the concept-level analysis of the all-or-none concept model[3] (Restle, 1961). The reader wishing to review these analyses in greater detail is referred to Atkinson, Bower, and Crothers (1965, Chapters 2, 3).

The (P,R) and (R,P) analyses are less usual and require some comment. A (P,R) analysis consists of plotting data statistics on the P-level when data has been generated by the one-element R-level model. In other words the model implies the unit is the collection of M items related by a rule; the learning of this unit is governed by the R-level model; however, for analysis, the unit is broken into P-level subsequences — one for each item. Rather than present a simulation of data generated and analyzed in this way, the derivation of P-level statistics for arbitrary parameter values of the one-element R-level model will be presented. These derivations assume the anticipation procedure.

---

[3] Restle's model has a learning only on errors assumption; whereas, the one-element R-level model assumes learning is equally probable after a success or an error.

The (R,P) analysis is analogous to the (P,R) analysis except that data are examined on the R-level and the model which generates the data is a P-level model. In other words, several units (items in this case) are combined into a single unit and studied.

To undertake the comparison of these four possibilities a set of statistics was selected. These statistics were selected both because they are among those usually considered in applications of models to verbal learning data (cf. Bower, 1961) and because they reflect salient points to be made in the analysis. These statistics are the learning curve, probability of an error on trial $n+1$ given an error on trial n, probability of no more errors following an error on trial n, distribution, mean, and variance of the total errors T, distribution and mean of the trial of the last error L, and the probability of an error on trial n prior to the last error.

To avoid future confusion a word about the meaning of "trial" is in order. By a trial on a unit is meant any presentation of any member of that unit, and by the k^{th} trial on a unit is meant the k^{th} occurrence of members of the unit. To illustrate, consider the list on p. 11. The fifth trial of the P-level analysis would refer to an item's performance on the fifth cycle through the list, i.e., performance somewhere in the trial block 37-45 depending on when the item appears on its fifth cycle.

However the fifth R-level trial would occur somewhere midway in the second cycle of the list; i.e., the fifth R-level trial would refer to performance on the trial number of the fifth presentation of members of a category. This event would be constrained by the anticipation procedure to take place midway in the second cycle through the list.

Before presenting the results of the analysis, a further word about notation is needed. Define x_n to be the error-success random variable as follows:

$$x_n = \begin{cases} 1 & \text{if error on } n^{th} \text{ trial of unit} \\ 0 & \text{if success on trial } n . \end{cases}$$

Let T be the total error random variable ($T = k$ means k errors made on a particular unit), and let L represent the trial number of the last error. Then the statistics chosen for the comparison are the following, for $n \geq 1$ and $k \geq 0$:

1. $Pr(x_n = 1)$,
2. $Pr(x_{n+1} = 1 | x_n = 1)$,
3. $b_n = Pr(\text{no more errors following an error on } n)$,
4. $Pr(T = k)$, $E(T)$, $Var(T)$,
5. $Pr(L = k)$, $E(L)$,
6. $Pr(x_n = 1 | L > n)$.

The interesting comparisons of the four situations involve fixing the level of data analysis and varying the model. This is what is usually done in comparative studies of models (cf. Atkinson and Crothers, 1964). In Table 2.1 the (P,P) and (P,R) analyses are compared and in Table 2.2 the (R,P) and (R,R) analyses are considered. Appendix I illustrates typical derivations of equations presented in Tables 2.1 and 2.2. In these tables we shall refer to the parameters of the usual one-element model by c' and g'. c and g will be the parameters of the model analyzed on the inappropriate level, i.e., c' and g' for (P,P) and (R,R) and c,g for (P,R) and (R,P). We shall assume M items are paired to each response. Those readers not interested in pondering the tedious derivations in Appendix I may note that for $M = 1$, expressions derived for (P,R) and (R,P) should take the same form as those of (P,P) and (R,R) respectively.

Certain similarities in expressions under (P,P) and (P,R) are evident from the table. $\Pr(x_n = 1)$, $\Pr(T = k)$, and $\Pr(L = n)$ are geometric distributions for both (P,P) and (P,R). Also $\Pr(x_{n+1}=1|x_n=1)$, b_n, and $\Pr(x_n = 1|L > n)$ are constant over trials for both situations.

It is, however, immediately evident that a one-element model will not fit data statistics in (P,R). There are a number of ways to demonstrate this and one will be presented. Suppose that the one-element P-level model does fit data statistics in (P,R). Then, from Table 2.1, we have

$$\Pr^{(P,P)}(x_n = 1|L > n) = \Pr^{(P,R)}(x_n = 1|L > n) \ ,$$

which requires the functional identity

$$1-g' = 1-g$$

or

$$g' = g \ . \qquad (2.1)$$

Now equating expressions for $\Pr(x_n = 1)$ yields the identity

$$(1-g')(1-c')^{n-1} = \frac{(1-g)[1-(1-c)^M][(1-c)^M]^{n-1}}{Mc} \ ,$$

which, inserting (2.1), requires

$$(1-c')^{n-1} = \frac{[1-(1-c)^M]}{Mc}[(1-c)^M]^{n-1} \ . \qquad (2.2)$$

Table 2.1 Comparison of (P,P) and (P,R).

Statistic	(P,P) analysis	(P,R) analysis
1. $Pr(x_n = 1)$	$(1-g')(1-c')^{n-1}$	$\frac{(1-g)[1-(1-c)^M]}{Mc}[(1-c)^M]^{n-1}$
2. $Pr(x_{n+1}=1 \mid x_n=1)$	$(1-g')(1-c')$	$(1-g)(1-c)^M$
3. b_n	$\frac{c'}{1-g'(1-c')} = b'$	$\frac{1-(1-c)^M}{1-g(1-c)^M} = b$
4. i. $Pr(T = 0)$	$g'b'$	$1 - \frac{(1-g)b}{Mc}$
ii. $Pr(T = k)$ $(k > 0)$	$(1-g'b')(1-b')^{k-1}b'$	$\frac{(1-g)b^2}{Mc}(1-b)^{k-1}$
iii. $E(T)$	$\frac{1-g'}{c'}$	$\frac{1-g}{Mc}$
iv. $Var(T)$	$E(T)[\frac{2-b'}{b'} - E(T)]$	$E(T)[\frac{2-b}{b} - E(T)]$
5. i. $Pr(L = 0)$	$g'b'$	$1 - \frac{(1-g)b}{Mc}$
ii. $Pr(L = n)$ $(n > 0)$	$(1-g')(1-c')^{n-1}b'$	$\frac{(1-g)[1-(1-c)^M]}{Mc} b[(1-c)^M]^{n-1}$
iii. $E(L)$	$\frac{(1-g')b'}{c'^2}$	$\frac{(1-g)}{Mc[1-g(1-c)^M]}$
6. $Pr(x_n=1 \mid L > n)$	$(1-g')$	$(1-g)$.

(2.2) is satisfied only if $M = 1$, and $c' = c$, but then the R-level model would reduce to the P-level model. Thus, unless $M = 1$, the (P,R) analysis can not be fit by the one-element P-level model, i.e., the two models are not equivalent on the P-level.

Thus similarities in equation type exist between (P,P) and (P,R); however, the expressions are different functions of the parameters. After presenting the results of the (R,P) vs. (R,R) comparison, several other comparisons between (P,P) and (P,R) not depending on the choice of a particular model will be developed. Also in Chapter 3 a model involving both levels of learning will be presented, and the relative contribution of each sort of learning will be assessed.

In Table 2.2 the comparison between (R,R) and (R,P) is presented. The contrasts are more striking than for (P,P) vs. (P,R), so not all statistics will be presented in closed form. Again we are assuming a list of size M. Finally one further convention is needed. If N is an R-level trial we need the cycle number $K(N)$ of an item appearing on that trial. Since the K^{th} P-trial of an item is restricted to the R-trial interval $((K-1)M + 1, KM)$ we have

$$(2.3) \qquad K(N) = \max\{k: M(k-1) < N\} .$$

In cases where it is obvious we will denote $K(N)$ by K. Table 2.2 now follows.

Table 2.2 Comparison of (R,R) and (R,P)

Statistic	(R,R) analysis	(R,P) analysis
1. $\Pr(x_N = 1)$	$(1-g')(1-c')^{N-1}$	$(1-g)(1-c)^{K(N)-1}$
2. $\Pr(x_{N+1}=1 \mid x_N=1)$	$(1-g')(1-c')$	$\begin{cases} \frac{1}{M}[(1-c)(1-g)]+\frac{M-1}{M}(1-c)^K(1-g) & \text{if } N \text{ Mod } M = 0 \\ (1-c)^{K-1}(1-g) & \text{if } N \text{ Mod } M \neq 0. \end{cases}$
3. b_N	$\frac{c'}{1-g'(1-c')} = b'$	$b^*\{gb^*+\frac{1-g}{c}b^*[1-(1-c)^{K-1}]\}^{N-M(K(N)-1)}$ $\times \{gb^*+\frac{(1-g)}{c}b^*[1-(1-c)^K]\}^{M-1-(N-M(K-1))}$, where $b^*=\frac{c}{1-g(1-c)}$. This function increases with N.
4. i) $E(T)$	$\frac{1-g'}{c'}$	$\frac{M(1-g)}{c}$
ii) $\mathrm{Var}(T)$	$\frac{1-g'}{c'}[\frac{1-g'}{c'}(1-2c')+1]$	$\frac{M(1-g)}{c}[\frac{1-g}{c}(1-2c)+1]$
iii) $\Pr(T = 0)$	$\frac{g'c'}{[1-c'(1-g')]}$	$\{\frac{gc}{[1-g(1-c)]}\}^M$
iv) $\Pr(T = k)$ for $k > 0$	$(1-g'b')(1-b')^{k-1}b'$	Not obtainable in closed form by the writer.
5. i) $\Pr(L = 0)$	$\frac{g'c'}{[1-c'(1-g')]}$	$\{\frac{gc}{[1-g(1-c)]}\}^M$
ii) $\Pr(L = N)$ for $N > 0$	$(1-g')(1-c')^{N-1}b'$	$\begin{cases} g^{M-N}[(1-g)b^*]b^{*M-1} & N \leq M \\ (1-c)^{K-1}(1-g)b^*L^{*MK-N}_{K-1} \; L^{*N-KM-1}_K, \end{cases}$ where $b^*=\frac{c}{1-g(1-c)}$ and L^*_K is the probability an item has its last error on or before its k^{th} cycle.
iii) $E(L)$	$\frac{(1-g')}{c'[1-g'(1-c')]}$	NOT DERIVED
6. $\Pr(x_N=1 \mid L > N)$	$(1-g')$	NOT DERIVED

There are several striking contrasts between (R,P) and (R,R). $\Pr(x_N = 1)$ for (R,P) is flat in periods of M R-trials, i.e., $\Pr(x_N = 1)$ is equal to $(1-g)$ for the first M trials, $(1-c)(1-g)$ for the second M trials, and $(1-c)^2(1-g)$ for the third M trials, etc. In addition most other trial-dependent statistics take jumps on trials $kM + 1$ for $k = 0,1,2,\ldots$. Finally several statistics are not constant with trials for (R,P) but are for (R,R), e.g., $\Pr(x_{N+1} = 1 | x_N = 1)$.

The similarities between (R,P) and (R,R) are few. When they do exist, they derive from the fact that the learning of each item procedes independently. This is most strikingly seen in $\Pr(T = 0)$, $E(T)$, and $\mathrm{Var}(T)$.

In summary this section has illustrated that the choice of a level of data analysis can influence the appearance of data statistics in much the same way that a model, if valid, can influence these statistics. A second point is that a model not only generates predictions for statistics on the intended level of analysis, but it also generates predictions on any level. This fact suggests that analyses on several levels in an experiment might provide supporting evidence for the validity of a model. The next section presents some cross-level analyses not restricted by choice of model.

Model-Free Analyses of P- and R-level Learning

Next we discuss model-free methods for determining when some learning takes place at a higher level (more R- like) or a lower level (more P-like) than the level of data analysis. What is meant by "model-free" needs some clarification. We view performance, not learning. Thus some sort of theory (or model) must be assumed to infer learning from performance data. By model-free is meant that we are assuming only that a change in the degree of learning of a unit manifests itself in a corresponding change in probability of correct to all items in that unit (i.e., an operational definition of "learning on a level" is desired).

In this section we parallel the structure of the preceding section. First methods for determining when learning takes place at a higher level than level of analysis will be discussed, and then indications of when learning takes place at a lower level than the level of analysis will be developed. For ease of presentation we will present these results in the context of the P- and R-levels of the preceding section. It should be clear how to generalize these results to the case where more levels exist.

Now we consider methods of indicating when learning is at a higher level than analysis. Accordingly, consider the case where some learning takes place on the R-level. For simplicity suppose $M = 2$, i.e., pairs of related items are assigned the same response. Imagine the two P-level protocols for an item pair are lined up one above the other. Since, by assumption, a single learning event may have resulted in simultaneous learning of both items in the pair, the sequences should bear a relationship to each other. For example if the one-element R-level model held with $g = 0$, the pair of last error trials for the two protocols would differ by at most one trial. Thus, if S_1 and S_2 are the two stimuli, their protocols might look like the following:

S_1 1111111000000...

S_2 1111111100000... .

In general any tendency for R-level learning should produce "covariation" in protocol pairs of related items. Thus if Z^i is a statistic for the i^{th} protocol $\rho_{Z^1Z^2}$ should be non-zero.

To illustrate, let x_n^1 and x_n^2 be the error-success (1-0) random variables for S_1 and S_2. Suppose the one-element R-level model holds with c,g free. It is a simple calculation to derive $\rho_{x_n^1 x_n^2}$.

$$\rho_{x_n^1 x_n^2} = \frac{\operatorname{Cov}\{x_n^1 x_n^2\}}{S_{x_n^1} S_{x_n^2}} \, .$$

(2.4)
$$\operatorname{Cov}\{x_n^1 \, x_n^2\} = E(x_n^1 \, x_n^2) - E(x_n^1)E(x_n^2)$$
$$= \Pr(x_n^1 = 1, \, x_n^2 = 1) - \Pr(x_n^1 = 1) \Pr(x_n^2 = 1).$$

Taking the two possible orders of presentation of S_1 and S_2 on P-trial n into consideration we have

$$\Pr(x_n^1 = 1, \, x_n^2 = 1) = \frac{(2-c)(1-g)}{2} \Pr(x_n^2 = 1) \, .$$

Since

(2.5)
$$S^2_{x_n^1} = S^2_{x_n^2} = \Pr(x_n = 1) \, ,$$

and

(2.6)
$$E(x_n^1) = E(x_n^2) = \Pr(x_n = 1) \, ,$$

we have

(2.7)
$$\rho_{x_n^1 x_n^2} = \frac{(2-c)(1-g)}{2} \, [1-(1-c)^{2(n-1)}] \, .$$

The function starts at a value 0 on trial 1 and increases exponentially to an asymptote of $\frac{(2-c)(1-g)}{2}$.

Of course the sample variance of the statistic $\rho_{x_n^1 x_n^2}$ would increase with n as fewer errors are made. Although not presented here, this sampling variance could be calculated from the model. Thus the properties, including power, of a test of zero $\rho_{x_n^1 x_n^2}$ could be established. In general, $\rho_{x_n^1 x_n^2}$ should be fairly simple to compute for any R-level model (or even a model which allows both P- and R-level learning such as the one-element multi-level model presented in the next chapter) provided the model is in any way tractable.

Other statistics could have been chosen for a correlation analysis. Several experimenters have empirically correlated total errors in an effort to ascertain relationships among units in the learning phase (Suppes, et al., 1962; Crothers and Suppes, in press). For example in Chapter 5 of the Crothers and Suppes' book, subjects were required to make multiple-choice grammatical ending responses to Russian nouns. Several grammatical classes served as the "concepts" to be learned. Various theoretical schemes for predicting the course of learning were presented. They were assessed on their ability to account for the pattern of pair-wise part correlations of total errors to the various concept classes.

This writer would suggest that matrices of part correlations of statistics such as total errors or trial of the last error could be used often as a device for checking whether some learning is taking place on a higher level than analysis. This procedure can be illustrated by an unpublished experiment by D. R. Rumelhart and the writer. Only the analysis relevant to the correlation method will be presented now.

In this study college-age subjects learned to pair 24 highly structured stimuli to 6 response classes by the anticipation procedure. The S-R pairs (which were consonant letters) had the following structure:

	Stimulus	Response
1.	ACE	1
2.	ACF	1
3.	ADE	1
4.	ADF	1
5.	BCE	2
6.	BCF	2
7.	BDE	2
8.	BDF	2
9.	IGK	3
10.	IGL	3
11.	JGK	3
12.	JGL	3
13.	IHK	4
14.	IHL	4
15.	JHK	4
16.	JHL	4
17.	OQM	5
18.	ORM	5
19.	PQM	5
20.	PRM	5
21.	OQN	6
22.	ORN	6
23.	PQN	6
24.	PRN	6

It should be noted that successive groups of four stimuli have a common letter and are paired to the same response.

The learning data appear very complicated and their analysis is only partially complete at this writing. It appears that learning has taken place on several levels in the experiment. This fact was tested by correlating trials of the last error to items in each group of four. Without presenting the details of this analysis here, it demonstrated a highly significant tendency for items in a 4-unit to have similar last error trials. By subtracting each subject's mean trial of the last error from each of his 24 items, a control for individual differences was attempted, i.e., the data for the analysis were of the form

$$L^{ij} - \frac{1}{24}\sum_{i=1}^{24} L^{ij} ,$$

where L^{ij} is the last error trial for item i subject j. More will be said about this experiment in the next section of this chapter and in Chapter 6.

Thus far we have considered in some detail the implications of learning on a level higher than the level of analysis. The conclusion was to compute correlations of various statistics on the units of the level of analysis. Any significant non zero correlation could be interpreted as a possible indication of higher level learning.

Next we return to the question of the implications of learning at a lower level than data analysis. The answers here are quite simple. Consider the R-level analysis of P-level data. It is a property of P-level learning, regardless of the model, that every M trials there will be a jump in the learning curve, i.e., $Pr(x_N = 1)$ will be flat in periods of M trials. This result comes directly from the anticipation procedure and the assumption of P-level learning which implies that items are learned independently.

In addition a statistic such as total errors is easy to work with. Regardless of the model we have

$$\mathrm{Var}(T_R) = M\ \mathrm{Var}(T_P')$$

where T_R is the total error random variable for the R-level and T_P' is the total error random variable for an item. This result comes from the independence assumption.

To illustrate these methods consider the experiment by Rumelhart and the writer discussed on pp. 24-26. $\Pr(x_N = 1)$ is plotted in Fig. 2.1 for the R-level analysis $(M = 4)$. A definite tendency for $\Pr(x_N = 1)$ to drop within a cycle indicates some R-level learning. The sizable jumps in $\Pr(x_N = 1)$ between cycle 2 and cycle 3 might indicate some P-level learning.

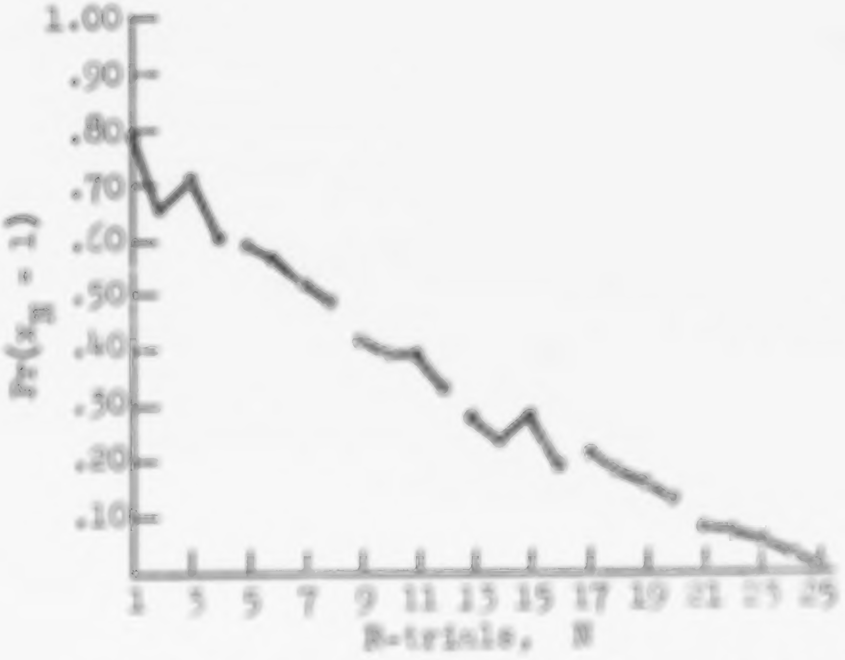


Fig. 2.1. R-level Learning Curve for the list depicted on p. 25 $(M = 4)$.

The R-level learning curve is also used to show some R-level and some P-level learning for other experiments in Chapter 6 (p. 134, Fig. 6.12).

In addition to the learning curve and $Var(T_R)$, $Pr(x_{n+1} = 1 | x_n = 1)$ should jump on trials $kM + 1$, $k = 1,2,\ldots,$ and the stationarity curve should rise over trials. Of course this latter feature could be accounted for by other P-level models such as the two-element model (Suppes and Ginsberg, 1963).

Conclusion

In this chapter we have discussed some of the implications of learning on various levels. Two methods of inferring level of learning have been developed, though not exhaustively. The first is to assume a model and then derive statistics for analyses on several levels. Inferences can then be made on the basis of the fit of the model to the data. The second method involves considering the general properties of the assumption of learning at a certain level. These properties, which depend on the mode of item presentation, suggest several statistical analyses, e.g., $\rho_{x_n^1 x_n^2}$. This chapter will have served its purpose if it convinces the reader that valuable inferences can be made from analyses of data on several levels.

CHAPTER 3

THE ALL-OR-NONE MULTI-LEVEL MODEL

The derivations and results of the previous chapter dealt mainly with the case where learning was assumed to take place on either the P-level or the R-level but not both. In cases where there would be any question of which level learning takes place at the more likely possibility would seem to be some learning on both levels. The question then arises as to whether extant verbal learning models, such as the one-element model, can naturally be generalized to allow for learning on several levels simultaneously. In this chapter a simple generalization of the one-element model to allow for such simultaneous learning is developed. In the next chapter a framework is proposed for axiomatizing other multi-level models.

The model to be developed in this chapter (the all-or-none multi-level model) is intended to be a simple and natural extension of the one-element P- and R-level models. It is not intended to represent a theoretical stand on the issue of how paired-associate learning takes place. So, rather than regarding this model as an addition to the crowded literature on paired-associate models, it should be regarded as an exercise in the synthesis of extant models.

Axioms for the Model

In the development to follow we will assume that subjects are learning a list with a structure similar to the list on p. 11, Chapter 2. In general, we assume that the list consists of J groups of M stimuli, where the members of any group are mutually related and each paired with the same response. By the stimuli in a group being related is meant that there is some common rule or common structure to all the stimuli connected to any particular response. Thus in the previously mentioned list the three rules are respectively: mathematicians are ones, chess players are twos, and football players are threes. No particular presentation schedule is assumed, but the model will be axiomatized under the assumption that on any presentation of a member of the list the subject first gives a response and then receives a paired presentation of the stimulus and its correct response (i.e., any particular presentation is like a particular presentation for the anticipation procedure.)

We wish to generalize the one-element model to allow for the possibility of learning the rule on any presentation of a relevant S-R pair and, in addition, to allow learning of that particular S-R pair if the rule is not learned. Accordingly, we will define an unlearned state U, an instance (paired-associate) state, P, and a rule-learned state R. We require that each of the M' items be in one and only one of these states on any trial. Transitions among these states are possible only when an item is presented, and the probabilities of these transitions do not depend on the past history of presentations and responses but only on the current state of the presented item.

The major departure from usual models is the assumption that if any item makes a transition to the R-state all other items on that trial move to the R-state. Thus an item's state may change when it is not presented. Finally performance (probability of a correct response) is assumed to be at a level g in state U and at a level 1 in states P and R.

More formally, let N index presentations of items in a block of M (i.e., R-trials); axioms for the all-or-none multi-level model are as follows:

1. Each of the M items is represented as being in exactly one of three states on any trial N. The states are an unlearned state U, an instance learned state P, and a rule learned state R.

2. All items start in state U, i.e., all items are in state U on R-trial N = 1.

3. When an item is presented it can change its state, and the probabilities governing these changes depend only on the current state of the presented item and not on the states of any of the other M-1 items, the past states of any of the M items, or the trial number.

   The assumptions about transitions to new states for a presented item are exhibited in the following stochastic matrix. The ij^th^ term in the matrix is the probability a presented item in state i will reside in state j on the next R-trial $(i,j \in \{U,P,R\})$.

$$\begin{array}{ccc} & & \text{State of item on} \\ & & \text{R-trial after Presentation} \end{array}$$

$$(3.1)\qquad \begin{array}{cc} \text{state of} & R \\ \text{item on} & P \\ \text{trial of} & \\ \text{presentation} & U \end{array} \begin{array}{c} \begin{array}{ccc} R & P & U \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ c & 1-c & 0 \\ r & p & 1-r-p \end{bmatrix} \end{array}.$$

4. If on any trial the presented item makes a transition to state R, the other M-1 items immediately make a transition to state R so that on all R-trials after this event all M items are in state R. Other than this possibility of transition, items not presented remain in their current states.

This axiom can be summarized by the following rule:

a. If the presented item is in state U, the other M-1 items all stay in their current states with probability 1-r and all move to state R with probability r;

and

b. If the presented item is in state P, the other M-1 items all stay in their current states with probability 1-c and all move to state R with probability c.

6. Let x_N be a random variable defined by

$$x_N = \begin{cases} 1 & \text{if error on R-trial } N \\ 0 & \text{if success on R-trial } N. \end{cases}$$

Then

$$Pr(x_N = 1) = \begin{cases} (1-g) & \text{if presented item in } U \\ 0 & \text{if presented item in } P \\ 0 & \text{if presented item in } R\,. \end{cases}$$

Theorems and Derivations

Some of the properties of this model will be presented in the theorems and derivations to follow. The first theorem shows that, under appropriate restrictions on the parameters of the all-or-none multi-level model, the one-element P-level model and the one-element R-level model are obtained.

Theorem 3.1

a. If $r = c = 0$ and $p \in (0,1)$, the all-or-none multi-level model is equivalent to a one-element P-level model.

b. If $p = 0$, $r = c$, and $r, c \in (0,1)$, the model is equivalent to a one-element R-level model.

Proof

a. If $r = c = 0$ and $p \in (0,1)$, items can change their state only when they are presented. Since all items start in state U, state R can never be obtained. Thus the restriction implies the all-or-none multi-level model can be summarized by the following stochastic matrix for each item. S_α, $\alpha = 1,2, \ldots, M$:

$$\begin{array}{c} \\ P_{n_\alpha} \\ U_{n_\alpha} \end{array} \begin{array}{c} \begin{array}{cc} P_{n_\alpha+1} & U_{n_\alpha+1} \end{array} \\ \left[\begin{array}{cc} 1 & 0 \\ p & 1-p \end{array}\right] \end{array} \qquad \begin{array}{c} \Pr(\text{correct}\,|\,\text{row state}) \\ \left[\begin{array}{c} 1 \\ g \end{array}\right] \end{array},$$

where n_α indexes presentations of item S_α (i.e., P-level trials on any item). This is the one-element P-level model.

b. If $p = 0$, $r = c$, $r, c \in (0,1)$, all M items start in U and any presentation results in a transition of all the items to state R with probability r. If N indexes R-level trials, the following stochastic matrix for all M items can be derived:

$$\begin{array}{c} \\ R_N \\ U_N \end{array} \begin{array}{c} \begin{array}{cc} R_{N+1} & U_{N+1} \end{array} \\ \left[\begin{array}{cc} 1 & 0 \\ r & 1-r \end{array}\right] \end{array} \qquad \begin{array}{c} \Pr(\text{correct}\,|\,\text{row state}) \\ \left[\begin{array}{c} 1 \\ g \end{array}\right] \end{array}.$$

This is the matrix for the one-element R-level model ||

Some additional notation will facilitate the statement of the next theorem. Suppose the M items in a block are ordered: $S_1, S_2, \ldots, S_M$. For each R-trial N define the state variable for the block, $\vec{T}_N$, to be $\vec{T}_N = (T_{1,N}, T_{2,N}, \ldots, T_{M,N})$, where $T_{k,N}$ is the state of item S_k

on trial N for $k = 1,2,\ldots,M$. The preceding axioms for the all-or-none multi-level model could easily be written in terms of the random variable $\vec{T}_N$, but this will not be done here (see p. 45, Chapter 4).

The next property of the model to be developed has implications for experiments involving post-learning transfer. Suppose that following an initial learning phase subjects are asked to make "best guess" responses to new stimuli. Suppose further that the new stimuli are constructed similar to stimuli in one of the blocks of M items, i.e., the new stimuli share a relationship or a rule with the other M stimuli in the block. It is a consequence of the following theorem that the more initial training trials on the block of M related items, the higher is the probability of the appropriate transfer response to these new items.

Theorem 3.2

If $r,p,c \in (0,1)$ and N indexes R-trials on a block of M stimuli, then

$$\lim_{N \to \infty} \Pr(\vec{T}_N = (R,R,\ldots,R)) = 1 .$$

Proof

The theorem follows from the fact that state $\vec{T} = (R,R,\ldots,R)$ is an absorbing state. Let $\theta = \min(c,r)$. By hypothesis $0 < \theta$. Since, on any trial N, the block of M items has either probability r or c of moving into state $(R,R,\ldots,R)$, we have

$$\Pr(\vec{T}_N = (R,R,\ldots,R)) \geq 1-(1-\theta)^{N-1} ,$$

hence

$$1 \geq \lim_{N \to \infty} \Pr(\vec{T}_N = (R,R,\ldots,R)) \geq \lim_{N \to \infty} [1-(1-\theta)^{N}] = 1 .$$

The above inequality implies

$$\lim_{N \to \infty} \Pr(\vec{T}_N = (R,R,\ldots,R)) = 1 \quad ||$$

The next Theorem and Lemma imply that the order in which items are presented does not affect the probabilities of being in the various states.

Theorem 3.3

Suppose S_i and S_j are presented on R-trials N and $N+1$, $i,j = 1,2,\ldots,M$. Then, for all possible states $\vec{t}, \vec{t'}$ of the set of M items,

$$\Pr(\vec{T}_{N+2} = \vec{t'} \mid \vec{T}_N = \vec{t}, S_{i,N}, S_{j,N+1})$$

$$= \Pr(\vec{T}_{N+2} = \vec{t'} \mid \vec{T}_N = \vec{t}, S_{j,N}, S_{i,N+1}) .$$

Proof

The apparatus necessary for a completely rigorous proof of this Theorem will not be developed until the next chapter. What follows is an outline of the main ideas in the proof. If $\vec{T}_N = (R,R,\ldots,R)$, the result is immediate, so assume $\vec{T}_N \neq (R,R,\ldots,R)$. Either $\vec{T}_{N+2} = (R,R,\ldots,R)$ or it does not. If $\vec{T}_{N+2} = (R,R,\ldots,R)$, then commutativity follows by noting, for all real numbers, a, b,

$$a + (1-a)b = b + (1-b)a .$$

Using this fact with $a = r$, $b = c$ establishes the result for $T_{N+2} = (R,R,\ldots,R)$.

If $\vec{T}_{N+2} \neq (R,R,\ldots,R)$, then a presentation of S_i can affect only the state of item S_i (similarily for S_j). Since these effects are independent, the order of appearance of S_i and S_j does not matter ||

The preceding theorem will receive more attention in the next chapter (p. 47). Next we state a lemma which provides a strong test for the all-or-none multi-level model.

Lemma 3.1

Suppose in the first N R-trials k_i S_i presentations are to be made, where $i = 1,2,\ldots,M$ and

$$\sum_{i=1}^{M} k_i = N .$$

Then the order in which these stimuli are presented does not affect the probability of beingin the various states on $N + 1$.

Proof

The lemma follows by repeated application of pairwise commutativity established in Theorem 3.3 ||

The preceding theorem and lemma provide both a strong test for the all-or-none multi-level model as well as a considerable reduction in the complexity of derivations from the model under certain presentation schedules. These points will be brought out in more detail in Chapter 5 (p. 100) where an additional analysis of the model (in terms of the framework to be developed in the next chapter) is presented.

Derivations for the Anticipation Procedure

The model can also be used to provide a synthesis for the results of the preceding chapter. Under the assumption that $r = c$ (rule learning is equi-probable from both the P and U states) the multi-level model reduces to a model that postulates two simultaneous all-or-none processes: one for P-level learning and one for R-level learning.

In the next few pages statistics for both the P-level and R-level will be presented under the assumption of an anticipation presentation schedule.

It should also be clear that under suitable additional restriction of the parameters p and r, results relevant to the four possibilities in Tables 2.1 and 2.2 of Chapter 2 can be obtained. Table 3.1 indicates the parameter restrictions which yield the four possibilities analyzed in the previous chapter (based on Theorem 3.1).

Table 3.1

Conditions under which the All-or-None Multi-level Model Reduces to the Four Analyses of the Preceding Chapter (Tables 2.1,2.2).

Chapter 2 analysis	Restrictions on Multi-level parameters	Level of data analysis of Multi-level model
(P-ANALYSIS, P-MODEL)	$c = r = 0$	P
(P-ANALYSIS, R-MODEL)	$c = r, p = 0$	P
(R-ANALYSIS, P-MODEL)	$c = r = 0$	R
(R-ANALYSIS, R-MODEL)	$c = r, p = 0$	R .

Thus, a statistic derived for the multi-level model should reduce to its corresponding expression in Table 2.1 or Table 2.2 of the preceding chapter if the indicated parameter restrictions are made. Exceptions are when r appears in the denominator of an expression, e.g., $\Pr(x_n=1)$ for the P-level analysis (see Table 3.2).

Only $\Pr(x_N = 1)$ and $\Pr(x_{N+1} = 1 | x_N = 1)$ have been presented for the R-level analysis. Some of the other results cannot be obtained by this writer in closed form, and others seem much too cumbersome and uninformative to present. The results of the P- and R-level analysis of the restricted (r=c) multi-level model appear in Table 3.2 to follow. It should be reiterated that the anticipation procedure is assumed and that each group of related items has M members. For selected derivations of these statistics, the reader is referred to Appendix I. Finally $K(N)$ refers to the cycle number corresponding to R-trial N (Eq. 2.3).

The all-or-none multi-level model is an intermediate model to the P- and R-level models in the sense that it postulates both P- and R-level learning. It is of some interest to compare the analyses of Table 3.2. with those analyses of the P and R models in the last chapter (Tables 2.1 and 2.2, pp. 18,20, respectively).

The results in Table 3.2 for the P-level analysis bear a resemblance to the results for the P-level analysis of the one-element R-level model (Table 2.1). $Pr(x_n = 1)$ is a geometric function of n, and $Pr(T = k)$ and $Pr(L = n)$ are geometric distributions. Similarly $Pr(x_{n+1}=1|x_n=1)$, b_n, and $Pr(x_n = 1|L > n)$ are constants. Even with these similarities (which also hold for the usual one-element model) the multi-level model is an alternative to the P-level analysis of the one-element R-level model. This can be shown by comparing selected statistics in Table 3.2 with those of Table 2.1.

Denote by $\underline{R}$ the one-element R-level model and by $\underline{L}$ the multi-level model, assume both models are analyzed on the P-level. Denote the parameters of $\underline{R}$ by c', g' and those for $\underline{L}$ by p, r, g. Assume the models are equivalent. Then, by equating $Pr(x_n = 1|L > n)$, we have the functional identity

$$(3.2) \qquad g' = g .$$

Table 3.2. P- and B-level Analyses of the All-or-None Multi-level Model $(r = c)$.

Statistic	P-level Analysis	B-level Analysis
1. $Pr(x_n = 1)$	$\frac{(1-g)[1-(1-c)^N]}{Nc}[(1-g-c)(1-c)^{N-1}]^{n-1}$	$(1-g)(1-c)^{N-1}\left(\frac{(1-g-c)}{(1-c)}\right)^{f(N)-1}$
2. $Pr(x_{n+1}=1 \mid x_n=1)$	$(1-g)(1-g-c)(1-c)^{N-1}$	$\frac{1-g}{N}[(1-c-g)+(N-1)(1-c)(1-g)^{f(N)}]$ if n Mod $N = 0$ $(1-g)(1-c)(1-g)^{f(N)-1}$ if n Mod $N \neq 0$
3. b	$\frac{1-(1-g-c)(1-c)^{N-1}}{1-g(1-g-c)(1-c)^{N-1}}$	
4. i. $Pr(T = 0)$	$1 - \frac{(1-g)[1-(1-c)^N]}{Nc[1-g(1-g-c)(1-c)^{N-1}]}$	
ii. $Pr(T = k)$ $(k > 0)$	$b(1-b)^{k-1}\frac{(1-g)[1-(1-c)^N]}{Nc[1-g(1-c)^N]}$	
iii. $E(T)$	$\frac{(1-g)[1-(1-c)^N]}{Nc[1-g(1-c)^N]b}$	
iv. $Var(T)$	$E(T)[\frac{2-b}{b} - E(T)]$	
5. i. $Pr(L = 0)$	$1 - \frac{(1-g)[1-(1-c)^N]}{Nc[1-g(1-g-c)(1-c)^{N-1}]}$	
ii. $Pr(L = k)$ $(k > 0)$	$\frac{(1-g)[1-(1-c)^N][1-(1-g-c)(1-c)^{N-1}][(1-g-c)(1-c)^{N-1}]^{k-1}}{Nc[1-g(1-g-c)(1-c)^{N-1}]}$	
iii. $E(L)$	$\frac{(1-g)[1-(1-c)^N]}{Nc[1-g(1-g-c)(1-c)^{N-1}][1-(1-g-c)(1-c)^{N-1}]}$	
6. $Pr(x_n=1 \mid L > n)$	$(1-g)$.	

Comparing $\Pr(x_{n+1} = 1|x_n = 1)$ for $\underline{R}$ and $\underline{L}$ yields the identity

$$(1-g')(1-c')^M = (1-g)(1-p-c)(1-c)^{M-1} ,$$

which, inserting 3.2, reduces to

$$(1-c')^M = (1-p-c)(1-c)^{M-1} . \tag{3.3}$$

Now comparing $\Pr(x_n = 1)$ for both models yields the identity

$$\frac{(1-g')[1-(1-c')^M]}{Mc'} [(1-c')^M]^{n-1}$$

$$= \frac{(1-g)[1-(1-c)^M]}{Mc} [(1-p-c)(1-c)^{M-1}]^{n-1} .$$

Substituting (3.2) and (3.3) yields

$$\frac{1-(1-c')^M}{c'} = \frac{1-(1-c)^M}{c} . \tag{3.4}$$

This last identity implies $c' = c$.

Now

$$c' = c$$

and

$$(1-c')^M = (1-p-c)(1-c)^{M-1}$$

only if $p = 0$, but in this case model $\underline{L}$ becomes model $\underline{R}$. Therefore we conclude that provided $p \neq 0$, the multi-level model is not equivalent to an R-level one-element model analyzed on the P-level.

Now we turn to a comparison of the R-level analysis of the multi-level model and the (R,P) analysis in Table 2.2. The two statistics presented in Table 3.2 for the multi-level model bear similarities to their counterparts of Table 2.2 in the preceding chapter. $\Pr(x_N = 1)$ jumps on trials kM+1 for k=1,2,... for both models, and $\Pr(x_{N+1}=1|x_N=1)$ is constant in successive blocks of M trials and jumps on trials kM+1.

The major difference between the two models is in $Pr(x_N = 1)$ between jump points (i.e., within a cycle). Within a cycle the one-element model $Pr(x_N = 1)$ is flat; whereas, for the multi-level model, it is geometric in shape. To see this, $Pr(x_N = 1)$ for the multi-level model is plotted, for $g = 1/5$, $r = 0.1$, $p = 0.5$, $M = 3$, in Fig. 3.1 below.

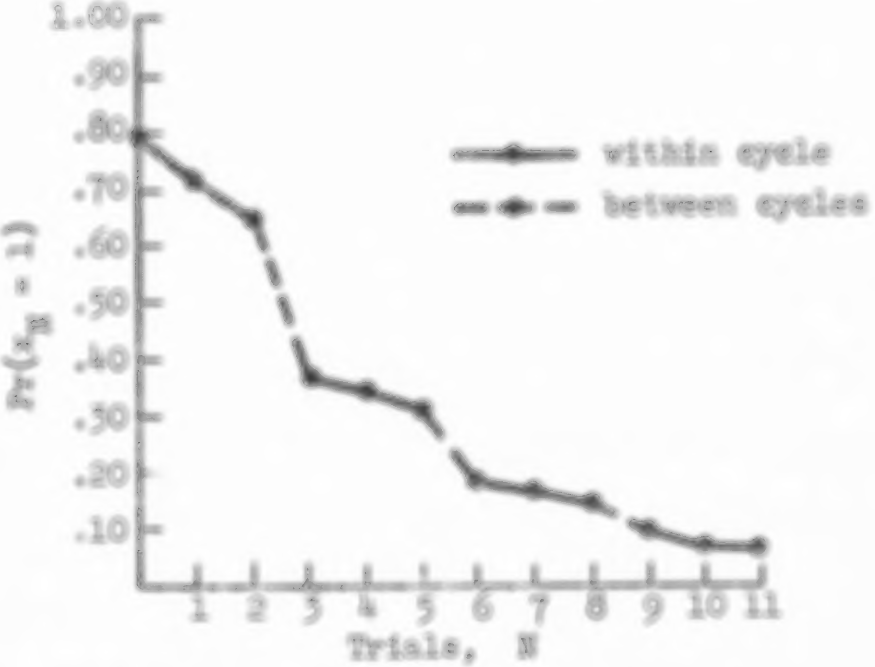


Fig. 3.1. R-level Learning Curve for All-or-None Multi-level Model ($g = 1/5$, $r = 0.1$, $p = 0.5$, $M = 3$).

Next we discuss correlation of item protocols for the multi-level model. Just as for the one-element R-level model, one would expect any two item protocols for related items to "co-vary." In the preceding chapter we introduced $\rho_{x_n^1 x_n^2}$ (Eq. 2.4) as a trial dependent measure of this co-variation (where x_n^i is the error-success random variable for item i). Assume $M=2$, then $\rho_{x_n^1 x_n^2}$ for the multi-level model ($r=c$) is

$$(3.5)\qquad \rho_{x_n^1 x_n^2} = \frac{\Pr(x_n^1 = 1,\ x_n^2 = 1) - \Pr(x_n^1 = 1)\,\Pr(x_n^2 = 1)}{\sqrt{\Pr(x_n^1 = 1)}\qquad\sqrt{\Pr(x_n^2 = 1)}}$$

$$= \Pr(x_n^1 = 1 \mid x_n^2 = 1) - \Pr(x_n = 1)$$

$$= \frac{(1-g)(2-r)}{2}\left[(1-g)^{n-1} - (1-g-r)(1-r)^{n-1}\right] .$$

$\rho_{x_n^1 x_n^2}$ is different for the one-element R-level model than for the multi-level model. For the one-element R-level model, $\rho_{x_n^1 x_n^2}$ starts at 0 for $n=1$ and increases to an asymptote of $\frac{(1-g)(2-r)}{2}$ as n increases. However $\rho_{x_n^1 x_n^2}$ for the multi-level model starts at 0, reaches a maximum for some $n > 0$, and then decreases to an asymptotic value of 0. This latter fact is true because, for large n, joint errors are only made to pairs of items not in state R and not both in state P. Given one item is not in P, the probability the other one is in P increases with n, i.e., the P-level process prior to the trial of transition into state R procedes independently for the two items.

An analysis of the general multi-level model with $r \neq c$ will be postponed until Chapter 5 (p. 98). This is because analysis of the model is greatly facilitated by a reformulation in terms of a general framework for analyzing multi-level models. The direction of this reformulation (alternative axiomatization) will be presented in the first section of the next chapter.

CHAPTER 4

GENERAL FRAMEWORK FOR MULTI-LEVEL MODELS

In the preceding chapter we have illustrated how a particular multi-level model might be developed. There is a property of this multi-level model that differentiates it from most other learning models. This property is that, under certain conditions, the $N-1$ items not presented on a trial change their states; whereas, under other conditions, only the presented item changes its state. In the axiom set f that model (Chapter 3, p. 31), it was awkward to formulate these properties. Thus the statement of Theorems 3.2 and 3.3 was greatly facilitated by the introduction of the random variable $\vec{T}_N$ (Chapter , p. 33), which keeps track of the states of all N items in a block. In addi on some further analyses of the model (Chapter 3, pp. 98-105) are greatly simplified by formulating the all-or-none multi-level model i terms of $\vec{T}_N$.

The organization of this chapter will be as follows. First the direction of reformulating the all-or-none multi-level model in terms of $\vec{T}_N$ will be indicated along with some of the advantages of this formulation over the formulation of the preceding chapter. This work will suggest a general framework within which many models that allow learning on several levels (or, equivalently, that allow items in a l st to mutually affect each other in the course of learning) can be axiom ized.

Before the framework is formalized, an indication of its intende scope will be presented. The scope of the framework will be presente by organizing the classes of models to which the framework can be applied under three headings. These headings will refer to three types of item dependencies (item interactions) permitted by the framework, and several examples of extant models embodying each type of dependency will be presented.

Finally the framework will be developed formally along with several theorems that can be applied to the analysis of any model axiomatized within the framework. The theorems fall into two classes. The first few theorems (Theorems 4.1, 4.2, 4.3) concern how to compute state probabilities and response probabilities for a model as a function of properties of the model and the presentation schedule. The latter theorems (Theorem 4.4, 4.5) concern how a model can be simplified along the lines of the particular dependencies it postulates, i.e., the framework will require that a model be stated in some generality and these theorems will concern how to reduce the generality in individual cases. The next chapter will present applications of the theorems to the analysis of the mixed model (Atkinson and Estes, 1963), the all-or-none multi-level model of Chapter 3, and a version of Restle's strategy-selection theory (Restle, 1962; 1964, Chapter 4).

To recapitulate the organization of this chapter, we will first reformulate the all-or-none multi-level model. This reformulation will suggest a general framework within which several models can be analyzed. Before presenting the formal aspects of the framework, an indication of the types of models which can be axiomatized in terms of the framework will be presented. Finally the framework will be developed along formal lines, i.e., definitions and theorems. Now we turn to the reformulation of the multi-level model.

Reformulation of the All-or-None Multi-level Model

Let us reconsider the all-or-none multi-level model. Suppose the M items in a block are ordered $S_1, S_2, \ldots, S_M$. Let $\vec{T}$ be a possible M-tuple of states for the M items, e.g., $\vec{T} = (R,R,\ldots,R)$. Define $\mathfrak{T}$ to be the set of all possible states of the M items. The axioms on p. 31 of the preceding chapter imply

$$\mathfrak{T} = \bigtimes_{n=1}^{M} \{U,F\} \cup \{(R,R,\ldots,R)\} \, , \tag{4.1}$$

where $\bigtimes_{n=1}^{M} \{U,F\}$ is the M-fold Cartesian product of the set $\{U,F\}$. $\mathfrak{T}$ has $2^M + 1$ members.

It is a property of the axioms for the model that, if the current state of the M items, $\vec{T}_N$, is known, and the presented item, $S_{i,N}$, is known, for $N=1,2,\ldots$, then the probabilities of being in the various 2^M+1 states in $\mathfrak{T}$ on trial $N+1$ are determined and are independent of past presentations, past states of the M items, and the trial index N. Suppose that the 2^M+1 members of $\mathfrak{T}$ are ordered, $\mathfrak{T} = (\vec{T}_1, \vec{T}_2, \ldots, \vec{T}_{2^M+1})$, then it is convenient to summarize the preceding remark by noting that the model implies that each of the items, S_i, has an associated set of transition probabilities from $\mathfrak{T}$ to $\mathfrak{T}$, where, for $i=1,2,\ldots,M$ and for all $\vec{T}, \vec{T}' \in \mathfrak{T}$,

$$\Pr(\vec{T}'_{N+1} \mid S_{i,N}, \vec{T}_N)$$

is determined (independent of N). It is desirable to represent these probabilities of transition from states in $\mathfrak{T}$ to states in $\mathfrak{T}$ by a stochastic matrix $\mathbb{P}_i$ for each item S_i, $i=1,2,\ldots,M$. Then, $\mathbb{P}_i$ is a $(2^M+1) \times (2^M+1)$ matrix of the probabilities of transition from states in $\mathcal{U}$ to states in $\tilde{\mathcal{U}}$ given S_i is presented. Thus, if item S_i is presented on some trial n, then the associated matrix $\mathbb{P}_i$ determines the probabilities of being in the various states, $\vec{T} \in \tilde{\mathcal{U}}$, given the current state of the set of M items. The $\mathbb{P}_i$ matrices are analogous to the stochastic matrices used to represent Markov learning models, e.g., the one-element P-level model has an associated stochastic matrix

$$\mathbb{P} = \begin{array}{cc} & \begin{array}{cc} P_{n+1} & U_{n+1} \end{array} \\ \begin{array}{c} P_n \\ U_n \end{array} & \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix} \end{array} .$$

As will be seen in Theorems 4.1, 4.2, 4.3, these matrices, $\mathbb{P}_i$, will be used to compute the probabilities of being in the various states given certain item presentation orders in much the same way as $\mathbb{P}$ is used to compute these probabilities for the one-element P-level model. The major difference in the two cases will be that the $\mathbb{P}_i$ matrices are used to compute the probabilities of being in various states of the entire list, whereas, $\mathbb{P}$ is used to compute the probability the presented item is in various states.

A reformulation of the all-or-none multi-level model can be accomplished in terms of the state space $\mathcal{U}$ and the M stochastic matrices, $\mathbb{P}_1, \cdots, \mathbb{P}_M$, defined in the preceding paragraph. An additional discussion of this reformulation is presented in Chapter 5, pp. 98-105. When the reformulation is done, it is much easier to state properties of the model than for the more conventional axiomitization of the preceding chapter. To illustrate, suppose $M=2$. Then $\tilde{\mathcal{U}} = \{(U,U),(U,P),(P,U),(P,P),(R,R)\}$. If the items are S_1 and S_2, we have

$$(4.2)\qquad \mathbb{P}_1 = \begin{array}{c} \\ (R,R) \\ (P,P) \\ (P,U) \\ (U,P) \\ (U,U) \end{array} \begin{array}{c} \begin{array}{ccccc} (R,R) & (P,P) & (P,U) & (U,P) & (U,U) \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 & 0 \\ c & 0 & 1-c & 0 & 0 \\ r & p & 0 & 1-r-p & 0 \\ r & 0 & p & 0 & 1-r-p \end{array}\right] \end{array},$$

and

$$(4.3)\qquad \mathbb{P}_2 = \begin{array}{c} \\ (R,R) \\ (P,P) \\ (P,U) \\ (U,P) \\ (U,U) \end{array} \begin{array}{c} \begin{array}{ccccc} (R,R) & (P,P) & (P,U) & (U,P) & (U,U) \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 & 0 \\ r & p & 1-r-p & 0 & 0 \\ c & 0 & 0 & 1-c & 0 \\ r & 0 & 0 & p & 1-r-p \end{array}\right] \end{array}.$$

Now, for example, to verify commutivity for the model with $M=2$ (Theorem 3.3, p.35), one merely needs to show that $\mathbb{P}_1 \cdot \mathbb{P}_2 = \mathbb{P}_2 \cdot \mathbb{P}_1$. The result is as follows:

$$(4.4)\qquad \mathbb{P}_1 \cdot \mathbb{P}_2 = \mathbb{P}_2 \cdot \mathbb{P}_1 = \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 1-(1-c)^2 & (1-c)^2 & 0 & 0 & 0 \\ c+(1-c)r & p(1-c) & (1-c)(1-r-p) & 0 & 0 \\ c+(1-c)r & p(1-c) & 0 & (1-c)(1-r-p) & 0 \\ r(2-r) & p^2 & p(1-r-p) & p(1-r-p) & (1-r-p)^2 \end{array}\right].$$

The basic idea of verifying that the $\mathbb{P}_i$ matrices commute provides the substance of Definition 4.7. Models which have this commuting property are much easier to work with than non-commuting models.

In addition to facilitating analysis of the model, the preceding formulation has another possible advantage. This advantage is that the model is stated in terms of the theoretical quantities $\mathcal{J}$ and the M matrices, which depend in no way on boundary conditions such as the presentation schedule or the level of data analysis. In other words, the stochastic process used to account for data in a particular experiment is not the model itself but a derivation from the model coupled with the particular presentation schedule and the level of data analysis. A model stated in this way can receive support from two sources: 1) its ability to make detailed predictions in a fixed situation (fixed schedule and level of analysis), and 2) its ability to account for the data in a number of different experiments in which both presentation schedule and level of data analysis vary. One illustration of the way boundary conditions are coupled with a model to derive a stochastic process for a fixed level of analysis is reported in Chapter 5 pp. 104-105. Theorems 4.2 and 4.5 are used for the all-or-none multi-level model (M=2). The anticipation presentation schedule is assumed and the level of data analysis is chosen as the error-success process on the first appearing item in a cycle, i.e., regardless of which of the two items is presented first on a cycle, the result of that trial is entered in the error-success protocol.

The preceding development is designed to preview the framework to be formalized in this chapter. It turns out that the framework is applicable to the analysis of many extant models which postulate item dependencies in the course of list learning. Before presenting the framework, the classes of models which can be axiomitized in terms of the framework will be organized around the types of item dependencies they postulate. This digression into other models has several motivations. First it is designed to show that the framework to be developed has wide applicability to extant models. Second it is the writer's feeling that more and more of the recent mathematical learning models are embodying some item dependencies in their assumptions (e.g., memory models). Thus it is becoming less and less often that models assume the learning of S-R pairs proceeds independently. It appears that one consequence of this tendency is that some methods of model analysis other than the traditional P-level analysis for the anticipation procedure are in order. With the knowledge that the case for this trend in mathematical learning theory can be made only by weight of evidence, we turn to this task.

If learning a list is presumed to take place on the P-level (level of individual items), then it is convenient to view each separate subject-item error-success (1-0) protocol as a sample path from some stochastic process whose sample space consists of all strings of 1s and 0s (cf. Atkinson, Bower, and Crothers, 1965, p. 82-83). If, on the other hand, the assumption of subject-item independence seems unrealistic, then this analysis is, at best, only approximately correct. A survey of some of the literature on mathematical learning models reveals that there are at least three distinct types of item dependencies postulated by models. This section presents a discussion of these three theoretical types of item dependencies, and then the framework, which is designed to incorporate the possibility of all three, is formalized.

The first type of item dependency postulated by some models is that response probabilities for a presented item may not depend solely on the state of the presented item, but also on the states of the other items in the list. The mixed model of Atkinson and Estes (1963) provides an example. In this model transitions among states for the presented items are independent of the states of unpresented items, but response probability to items in the unlearned state is determined by the states of all items in the list. The work of Friedman is related to the mixed model (Friedman and Gelfand, 1964; Friedman et al., 1966). In the Friedman, et al., paper, a three state Markov learning model on stimulus patterns is postulated, and a number of complex response rules involving stimulus components are developed.

Ruskin (unpublished doctoral dissertation) has analyzed the learning of concept stimuli composed of three two-valued dimensions in terms of models which assume that learning proceeds independently for each item, but that response probabilities to items in unlearned states depend on the states of all items in the list. He has had some success in accounting for differential numbers of errors to each stimulus in such problems.

The second type of item dependency postulated by models is that the state of an item can change on trials when it is not presented. The concept learning model of Restle (1961) fits into this category. Strictly speaking this hypothesis model has the property that the states of each item may or may not change when a new hypothesis is sampled. The usual all-or-none two-state model presented by Restle (1961) and also by Bower and Trabasso (1964) represents the process of concept learning in a much more simplified manner than their theory implies. They accomplish this by lumping the states of certain Markov chains implied by the theory. Even this simplified model has the property that items not presented can shift to the learned state.

More recently Restle (1962, 1964) has proposed a strategy-selection theory for paired-associate learning. The theory supposes that two similar items requiring dissimilar responses may become confused. Confusion is represented in the theory by certain mnemonical devices or strategies which the subject might use to retrieve an S-R pair from memory, i.e., if two S-R pairs in a list were AB-1, AC-2, then the strategy A-1 would result in confusion between AB and AC. It is a consequence of Restle's theory that an unpresented item, say AC in the above miniature list, can change its state when another item, AB, is presented. In Chapter 5, pp.108, Restle's model will be analyzed in detail using the framework to be developed in this chapter.

The all-or-none multi-level model presented in the previous chapter is another example of a model that allows states of items to change on trials when they are not presented. An additional analysis of this model in terms of the framework will be given in Chapter 5, p. 98.

A fourth example is the trial-dependent-forgetting model (T.D.F. model) of Atkinson and Crothers (1964) and Calfee and Atkinson (1965). In this model an item in a short term memory state can be bumped into a forgotten state as a consequence of the presentation of another unlearned item.

The third type of inter-item dependency postulated by models is that the state of a particular unpresented item can influence the transition probabilities for the presented item as well as other unpresented items. One example of this dependency is the Buffer Models of Atkinson and Shiffrin (1965). In these models the probability that an item will enter the short term memory buffer depends on the number of other items already in the buffer. Similarly whether or not an item in the buffer is dropped on a certain trial depends on how many other items are in the buffer. In most applications, however, the buffer is assumed to be full.

A second example is the two-person game situation discussed in Suppes and Atkinson (1960). Player A can be in response state A_1 or A_2, and the transition probabilities depend on the response state of player B in the sense that the states of both players determine the payoff probabilities, and the payoff determines, in turn, the transition probabilities.

A third example comes from a slight generalization of the all-or-none multi-level model presented in the last chapter. Suppose the probability of rule learning when all items are in U is r, but the probability of rule learning when any item is in P is $c \neq r$. Then a presented item in state U would have rule learning parameter r or c depending on the states of other items in the list.

In each of the examples presented, the probability of a response A_i to a presented item S_j on trial n, $Pr(A_{i,n}|S_{j,n})$, depends not only on the number of previous presentations of the item but also in some way on the number and positioning of presentations of items other than S_j. This seems to suggest that a useful testing ground for models embodying item dependencies is in experiments where the presentation orders are manipulated and predictions of the probabilities of various responses are made. It seems to this writer that the ability of a model to account for various patterns of response probability as a function of controlled presentation orders is every bit as strong a test of a model as its ability to account for subject-item error-success protocols in an anticipation procedure experiment. Of course the preceding remark assumes that the model makes differential predictions of response probability as a function of presentation order.

This presentation order approach to testing models which imply item dependencies has already been used by many, e.g., the miniature RTT paired-associate experiments (Estes, Hopkins, Crothers, 1960; Izawa, 1965; Young, unpublished doctoral dissertation), the work on optimization (Suppes, 1964; Crothers, 1965; Groen and Atkinson, in press); and work with memory models for paired-associate learning (Greeno, 1966; Atkinson and Shiffrin, 1965; Bjork, unpublished doctoral dissertation).

The Framework

A. History of Major Ideas

In this section a framework providing a possible synthesis of models which permit any of these three types of dependencies is developed. A number of general theorems for predicting state probabilities and response probabilities as a function of presentation sequence will be presented. The main theoretical quantity in the framework will be the state of the entire list rather than the more usual state of an item. The state of the list will be represented by a vector of states of the items in the list. Each item in the list will be characterized by a matrix of transition probabilities from states of the list to states of the list. A matrix associated with an item will be effective whenever that item is presented on a trial, i.e., to compute state probabilities on trial $N+1$ one applies the matrix operator associated with the item presented on trial N to the vector of probabilities of being in the various states of the list on trial N. A trial is defined to be the presentation of an item for a response, followed immediately by the item paired with its correct response, i.e., a trial here is equivalent to a usual anticipation procedure trial.

One precursor to the idea of defining a state of a list in terms of the states of the items in the list is found in Estes (1959). In developing general properties of component and pattern models, Estes suggests that one could define a state for a one-element paired-associate model in terms of the number of unlearned items in the list. The derivation on p. 36 of his chapter assumes the anticipation procedure. He shows how one can derive the probability of a correct response at the beginning of cycle n from a matrix whose states are the number of unlearned items, i.e., if the list has M stimuli, the states are 0,1,...,M. Estes' idea for treating the state of the list for the one-element model is generalized in this paper to apply to any model in the framework (Theorem 4.5).

A second example of the idea of combining states of various items into a single state is found in Atkinson and Estes (1963). In section 5.2 of their chapter on stimulus sampling theory, they develop the mixed model for a two item list. The items, ab and ac, are assumed to be either in an unlearned state U or a learned state L. They develop the theory for a four-state process with states (U,U), (U,L), (L,U), and (L,L), where the first position refers to the state of item ab and the second to item ac. More will be said about this work in Chapter 5, p. 91.

The idea that presentations of different items can be represented by different sets of transition probabilities among states of the entire list of items has been, in part, adopted by Restle (1962, 1964). His strategy-selection theory of paired-associate learning assumes that items in a list can be confused in the course of learning. This confusion results in a discrimination problem which is solved by discarding strategies that confuse items requiring dissimilar responses. Restle does not allow for different items to have different transition probabilities in his applications of strategy-selection theory (cf. Restle, 1964, Sec. 5, pp. 132-144); however, he points out that his applications are at best an approximation (Restle, 1964, pp. 168-171). In the final pages of the chapter, Restle suggests the direction necessary to take in order to square the models he uses with his theory. It is these suggestions of his, rather than his original model, that resemble certain developments in this chapter. A more detailed analysis of strategy-selection theory will be presented in Chapter 5 (pp.108) of this paper.

B. Definition of Model in Framework

In the development to follow each item in a list will be required to be in one of a number of finite states on any trial of the experiment. The generalization from usual formulations of models will be to allow for the possibility for some or all of the items which are not presented on a certain trial N to do any of the following: (1) affect response probabilities on trial N; (2) change their own states of conditioning on trial N; and (3) to affect transition probabilities of other items in the list on trial N.

Suppose a list of M S-R pairs (items), denoted by

$$\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$$

and a set of Q responses, denoted by

$$\mathcal{Q} = \{A_1, A_2, \ldots, A_Q\} .$$

We will adopt the idea of a state as a primitive notion in the framework. States U and L in the one-element model, the number of patterns connected to response A_1 in a two response pattern model, and U, P, and R in the one-element multi-level model are all states in the intended usage of "state" in the framework. In Definition 4.1 the notion of an item state space is presented. It should be noted that, since the item state space is an ordered set of states, it is possible for a particular state to appear more than once (with a different subscript) in the item state space.

Definition 4.1

By a state space, T_I, of an item is meant a finite ordered set of states

$$T_I = \{\tau_1, \tau_2, \ldots, \tau_L\} .$$

Examples of item state spaces are $\{U,L\}$ for the one-element model, $\{C_j\}_{j=0}^{N}$ for the N-element two-response pattern model, and $\{U,P,R\}$ for the all-or-none multi-level model presented in the preceding chapter. Next we formalize in Definition 4.2 the notion of the state space for a list of items.

Definition 4.2

By a state space $\mathcal{T}$ of a list of M items with item state space T_I is meant the M-fold Cartesian product

$$\mathfrak{T} = T_I \times \cdots \times T_I ,$$

where "x" is the Cartesian product of sets.

For the one-element P-level model with a list of M items

$$\mathfrak{T} = \{\vec{t} = (t_1, \ldots, t_M): \; t_i \in \{U,L\}, \; i = 1, 2, \ldots , M\} . \tag{4.5}$$

Thus if the item state space T_I has L states, the state space for a list with M items will have L^M members.

We will next define a model for the learning of a list (Definition 4.3). This definition will require that the stochastic process governing state-to-state transitions among $\mathfrak{T}$ be Markov in a certain sense. The Markov restriction is not thought to be too severe because in many non-Markov models the state space could be expanded to make the model satisfy the Markov condition. Disregarding the restriction of a finite state space for the moment, the identifiable state theory developed in Greeno and Steiner (1964, p.317) illustrates one way in which this expansion can be accomplished.

Although the restriction to a finite-state model and the Markov condition rule out certain models, like the linear model which requires an infinite state space to satisfy the Markov condition, generalization of the present approach to include these models should be possible.[4] We next present Definition 4.3.

Definition 4.3

Suppose $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$ is a list of M items with associated response set $\mathcal{A}$ of size Q. Then $\mathcal{M} = (\mathfrak{T}, \mathcal{P}, \mathcal{L})$ is a model for the learning of list $\mathcal{S}$ in case:

---

[4] For example response probability might be used to define a state, and operators for each item could be postulated.

1. There is an item state space

$$T_I = \{\tau_1, \tau_2, \cdots, \tau_L\}$$

such that

$$\mathcal{T} = T_I \times \cdots \times T_I$$

is a state space for the list $\mathcal{S}$.

2. $\mathcal{P}$ is a set of M $L^M \times L^M$ square matrices, $\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_M$, such that, for all $i, j = 1,2,\ldots,L^M$ and $\alpha = 1,2,\ldots,M$, the ij^{th} term of $\mathbb{P}_\alpha$, namely $\mathbb{P}_\alpha^{ij}$, is the probability of transition from state $\vec{t}_i$ of the list to state $\vec{t}_j$ on a trial when item S_α is presented. These transition probabilities depend only on i, j, α, and not on the trial index or preceding states of the list, i.e., for all trials N; stimuli $S_\alpha \in \mathcal{S}$; states of the list $\vec{U}_i, \vec{V}_j \in \mathcal{T}$, and past histories of presentations, responses, and states, h, we have

$$\mathbb{P}_\alpha^{ij} = \Pr(\vec{V}_{j,N+1} | S_{\alpha,N}, \vec{U}_{i,N}, h_N) .$$

3. $\mathcal{L}$ is a function which specifies, for each stimulus $S_i \in \mathcal{S}$, response $A_j \in \mathcal{A}$, and state of the list $\vec{t} \in \mathcal{T}$,

$$\Pr(A_{j,N} | S_{i,N}, \vec{t}_N)$$

independent of the trial number $N = 1, 2, \ldots$.

Naturally if $\mathcal{M}$ is a model we have, for all $\alpha = 1,2,\ldots,M$, $i = 1, 2,\ldots,L^M$,

$$\sum_{j=1}^{L^M} P_\alpha^{ij} = 1 ,$$

and, for all $j = 1, 2, \ldots, M$ and $\vec{t} \in \mathcal{T}$,

$$\sum_{i=1}^{Q} \Pr(A_{i,N} | S_{j,N}, \vec{t}_N) = 1 .$$

Consider, as an example, the one-element P-level model for a list of M items. The state space for this model is defined in Eq. 4.5, and the model specifies that transitions are possible only for the presented item. Denote by t^α the α^{th} component of $\vec{t}$, where $\vec{t} \in \mathcal{T}$. Suppose $S_\alpha \in \mathcal{S}$ is presented on some trial. Then the model implies for all $\vec{t}$, $\vec{w} \in \mathcal{T}$ with $t^\beta = w^\beta$ for $\beta \neq \alpha$,

$$(4.6) \qquad P_\alpha^{\vec{t}\vec{w}} = \begin{cases} 1 & \text{if } t^\alpha = L \text{ and } w^\alpha = L \\ 0 & \text{if } t^\alpha = L \text{ and } w^\alpha = U \\ c & \text{if } t^\alpha = U \text{ and } w^\alpha = L \\ 1-c & \text{if } t^\alpha = U \text{ and } w^\alpha = U , \end{cases}$$

and, if $\vec{t}$ and $\vec{w}$ do not agree in any coordinate β with $\beta \neq \alpha$,

$$P_\alpha^{\vec{t}\vec{w}} = 0 .$$

The response rule for the one-element model is generally stated in terms of a correct response and an incorrect response. Let A_α be the response associated with S_α and $\bar{A}_\alpha = \mathcal{A} - \{A_\alpha\}$. Then, for all $\alpha = 1, 2, \ldots, M$, $\vec{t} \in \mathcal{T}$ and $N = 1, 2, \ldots$,

$$(4.7) \qquad \Pr(A_{\alpha,N} | S_{\alpha,N}, \vec{t}_N) = \begin{cases} 1 & \text{if } t^\alpha = L \\ g & \text{if } t^\alpha = U , \end{cases}$$

and, of course,

$$\Pr(\bar{A}_{\alpha,N} | S_{\alpha,N}, \vec{t}_N) = \begin{cases} 0 & \text{if } t^\alpha = L \\ 1-g & \text{if } t^\alpha = U . \end{cases}$$

To further clarify these abstractions assume the one-element model for a list with $M=2$. The members of $\mathcal{T}$ are (U,U), (U,L), (L,U), and (L,L), where the first member is the state of S_1 and the second is the state of S_2. According to Definition 4.3 we have

$$(4.8) \qquad \mathbb{P}_1 = \begin{array}{c} \\ (L,L) \\ (L,U) \\ (U,L) \\ (U,U) \end{array} \begin{array}{c} \begin{array}{cccc} (L,L) & (L,U) & (U,L) & (U,U) \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ c & 0 & 1-c & 0 \\ 0 & c & 0 & 1-c \end{array}\right] \end{array} ,$$

and

$$(4.9) \qquad \mathbb{P}_2 = \begin{array}{c} \\ (L,L) \\ (L,U) \\ (U,L) \\ (U,U) \end{array} \begin{array}{c} \begin{array}{cccc} (L,L) & (L,U) & (U,L) & (U,U) \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & c & 1-c \end{array}\right] \end{array} .$$

One effect of the preceding definitions is to allow us to view a theory for list learning as a set of M matrices of transition probabilities among the L^M states of $\mathcal{T}$. The device of dealing with $\mathcal{P} = \{\mathbb{P}_1, \ldots, \mathbb{P}_M\}$ permits one to handle the possibility of simultaneous learning on various levels. To illustrate, suppose the all-or-none multi-level model is written for a list with $M=2$. Then the implied matrices $\mathbb{P}_1$ and $\mathbb{P}_2$ are given by Eqs. 4.2 and 4.3. The response rule for the all-or-none multi-level model specifies that response probabilities are completely determined by knowing the state of the presented item. In general, item dependencies implied by multi-level learning are recorded by their effect on $\vec{t} \in \mathcal{T}$.

C. Definitions and Theorems for Presentation Schedules

Now that the notion of a model in the framework has been formalized, we move to the task of stating theorems for computing state and response probabilities as a function of the presentation schedule. These theorems are motivated by the idea that a multi-level model can be tested by manipulating presentation sequences and predicting response probabilities as a function of this manipulation. Before stating the theorems of this section, one more definition is needed.

The formulation of the notion of a model in the preceding section did not include a specification of the probabilities of being in the various states of the list on trial one, i.e., Definition 4.3 did not include a start vector. In order to apply a model to a particular experiment a start vector must either be assumed by the model, or the probabilities of starting in the various states must be regarded as parameters of the model. The notion of a start vector is formalized in Definition 4.4.

Definition 4.4

By a start vector $\vec{p}_1$ for a model $\mathcal{M} = (\mathcal{J}, \mathcal{P}, \mathcal{L})$ is meant an L^M dimensional row vector of the probabilities of being in the L^M states in $\mathcal{J}$ at the start (trial one) of an experiment.

In general, denote by $\vec{p}_N$ the row vector of probabilities of being in the various L^M states on trial N of an experiment. $\vec{p}_N$ can be viewed as a random variable whose value depends on the start vector $\vec{p}_1$, the matrices $\mathbb{P}_1, \dots, \mathbb{P}_M$, and the presentation schedule.

Next we present Theorems 4.1, 4.2, and 4.3 which give general methods for computing $\vec{p}_N$ as well as response probabilities on trial N

under some frequently employed presentation schedules. The first theorem shows how to compute $\vec{p}_{N+1}$ for a fixed presentation sequence of stimuli for the first N trials. Theorem 4.1 should be regarded as a fairly obvious extension of a standard theorem in Markov chain theory. The theorem from Markov chain theory asserts that if $\mathbb{P}$ is the transition matrix for a finite state Markov chain, the probability of being in state j N trials after being in state i is given by the ij^{th} term of $\mathbb{P}^N$ (cf. Kemeny, Mirkil, Snell, and Thompson, 1959, p. 386). Theorem 4.1 is a special case of the analogous theorem for inhomogeneous finite state Markov chains (i.e., chains whose parameters are trial dependent).

Theorem 4.1

Suppose a list, $\mathcal{S}$, of M items, a model $\mathcal{M} = (\mathcal{T}, \mathcal{P}, \mathcal{L})$ with an associated start vector $\vec{p}_1$. Also suppose the presentation sequence $S_{\alpha_1}, S_{\alpha_2}, \ldots, S_{\alpha_N}$, for $S_{\alpha_i} \in \mathcal{S}$, $i = 1, 2, \ldots, N$, is administered for the first N trials. Then the row vector of probabilities of being in the various states of the list on trial $N + 1$ is given by

$$\vec{p}_{N+1} = \vec{p}_1 \prod_{i=1}^{N} \mathbb{P}_{\alpha_i} . \tag{4.10}$$

Proof

The proof proceeds by induction on N. Clearly for $N = 1$

$$\vec{p}_2 = \vec{p}_1 \mathbb{P}_{\alpha_1} .$$

Assume for $N - 1$ that

$$\vec{p}_N = \vec{p}_1 \mathbb{P}_{\alpha_1} \cdots \mathbb{P}_{\alpha_{N-1}}$$

Then the k^th term of $\vec{p}_{N+1}$ is given by

$$\begin{aligned}
\vec{p}_{N+1}^{(k)} &= \sum_{j=1}^{L^M} \vec{p}_N^{(j)} \, \mathbb{P}_{\alpha_N}^{jk} \\
&= \sum_{j=1}^{L^M} [\vec{p}_1 \, \mathbb{P}_{\alpha_1} \cdots \mathbb{P}_{\alpha_{N-1}}]^{(j)} \, \mathbb{P}_{\alpha_N}^{jk} \\
&= [\vec{p}_1 \, \mathbb{P}_{\alpha_1} \cdots \mathbb{P}_{\alpha_N}]^{(k)} .
\end{aligned}$$

Hence

$$\vec{p}_{N+1} = \vec{p}_1 \prod_{i=1}^{N} \mathbb{P}_{\alpha_i} \quad \|$$

Although the preceding theorem is not suited for hand computation of any but the most simple models with M and L small, it could provide a useful tool in computer simulation of more complex multi-level models.

Although this theorem and the ones to follow concern how to derive the probabilities of being in the various states given various presentation sequences, it is quite easy to use these results to get response probabilities. Suppose s_{N-1} is the sequence of presentations for the first N-1 trials; then, for all $S_i \in \mathcal{S}$, $A_j \in \mathcal{A}$, $\vec{t}_k \in \mathcal{T}$, and $N = 1, 2, \ldots$,

$$\begin{aligned}
\Pr(A_{j,N} | S_{i,N}, s_{N-1}) &= \sum_{k=1}^{L^M} \Pr(A_{j,N} | S_{i,N}, \vec{t}_{k,N}, s_{N-1}) \Pr(\vec{t}_{k,N} | s_{N-1}) \\
&= \sum_{k=1}^{L^M} \Pr(A_{j,N} | S_{i,N}, \vec{t}_{k,N}) \, \vec{p}_N^{(k)} .
\end{aligned} \tag{4.11}$$

The first term is given by $\mathcal{L}$ and the second is the k^th component of $\vec{p}_N$

calculated by Theorem 4.1. In later theorems $E(\vec{p}_N)$ will be computed under various presentation schedules. If J is a presentation schedule (see Definition 4.5), we have

$$(4.12) \qquad E_J(\Pr(A_{j,N}|S_{i,N})) = \sum_{k=1}^{L^M} \Pr(A_{j,N}|S_{i,N}, \vec{t}_{k,N}) E_J(\vec{p}_N^{\,(k)}) \quad .$$

Next we define the notion of a presentation schedule generator (p.s.g.) and state a lemma from Theorem 4.1 for finding $E(\vec{p}_N)$ under an arbitrary p.s.g.

Definition 4.5

Suppose a list of M items, $\mathcal{S}$. By a presentation schedule generator, J, is meant a rule which specifies the following probabilities:

1. For all presentations on the first trial, $S_\alpha \in \mathcal{S}$, $\Pr(S_{\alpha,1})$ is specified.
2. Let $\mathcal{S}_N$ denote $\underbrace{\mathcal{S} \times \ldots \times \mathcal{S}}_{N\text{-times}}$ and let h_N denote the history of the first N presentations and responses. Then, for all $s_N \in \mathcal{S}_N$, all histories h_N, and all $S_\alpha \in \mathcal{S}$,

$$\Pr(S_{\alpha,N+1}|s_N, h_N)$$

is specified.

Thus a p.s.g. is just a rule for determining the probability of any sequence of presentations through the first N trials (possibly contingent on the subject's responses) for $N = 1, 2, \ldots$.

Lemma 4.1

Suppose a p.s.g. J for a list of M items and a model $\mathcal{M}$ with associated start vector $\vec{p}_1$. Then the expected probabilities of being in the various states of the list on trial N (expectation with respect to J) are given by

$$(4.13) \qquad E_J(\vec{p}_N) = \sum_{s_{N-1} \in \mathcal{S}_{N-1}} (\vec{p}_1 \prod_{i=1}^{N-1} P_{\alpha_i}) Pr(S_{\alpha_1}, \dots, S_{\alpha_{N-1}})$$

where s_{N-1} is expressed as

$$S_{\alpha_1}, \dots, S_{\alpha_{N-1}} .$$

Proof

With Theorem 1 and the treatment of $\vec{p}_N$ as a random variable in mind, the lemma follows from the fact that if X and Y are discrete random variables

$$E(X) = \sum_y E(X | Y=y) Pr(Y=y) ,$$

where the sum is over y such that $Pr(Y=y) > 0$ ||

Next we introduce the notion of a "Bernoulli presentation schedule." A theorem is then stated for computing $E_B(\vec{p}_N)$ for a Bernoulli p.s.g. B. It turns out that for computational purposes it is useful to test a multi-level model with a Bernoulli presentation schedule. If this schedule is used, an expected operator or average transition matrix can be used to get state probabilities (Theorem 4.2), and a theorem which permits lumping of the average matrix (Theorem 4.5) under a further restriction greatly reduces the number of states in this matrix. These two theorems are used together to derive stochastic matrices for the all-or-none multi-level model and for Restle's strategy selection theory (Chapter 5, p. 103 and p. 115, respectively).

Definition 4.6

Suppose a list of M items $\mathcal{S}$. By a Bernoulli presentation schedule is meant a rule J which selects item $S_\alpha \in \mathcal{S}$ to be presented on Trial N with probability π_α independent of N and the previous item presentations and responses, $\alpha = 1, 2, \ldots, M$. That is, for all $\alpha = 1, 2, \ldots, M$, $N = 1, 2, \ldots,$

$$\Pr(S_{\alpha,N}) = \pi_\alpha$$

independent of N and the history h_{n-1}. Of course

$$\sum_{\alpha=1}^{M} \pi_\alpha = 1 .$$

Theorem 4.2

Suppose a list of M items, a model $\mathcal{M} = (\mathcal{T}, \mathcal{P}, \mathcal{L})$ with associated start vector $\vec{p}_1$ and a Bernoulli p.s.g. $B = \{\pi_1, \pi_2, \ldots, \pi_M\}$. Define

$$\mathbb{A} = \sum_{k=1}^{M} \pi_k \mathbb{P}_k$$

to be a matrix of "average" transition probabilities effective on any trial. Then

$$E_B(\vec{p}_{N+1}) = \vec{p}_1 \mathbb{A}^N . \tag{4.14}$$

Proof

The proof proceeds by induction on N. For $N = 2$,

$$\begin{aligned} E_B(\vec{p}_2) &= \pi_1 \vec{p}_1 \mathbb{P}_1 + \cdots + \pi_M \vec{p}_1 \mathbb{P}_M \\ &= \vec{p}_1 \sum_{k=1}^{M} \pi_k \mathbb{P}_k = \vec{p}_1 \mathbb{A}. \end{aligned}$$

Assume

$$E_B(\vec{p}_N) = \vec{p}_1 \mathbb{A}^{N-1} ,$$

Then for all strings $\alpha = (\alpha_1, \alpha_2, \dots , \alpha_N)$ of the first M integers we have

$$\begin{aligned} E_B(\vec{p}_{N+1}) &= \sum_{\alpha} \vec{p}_1 (\mathbb{P}_{\alpha_1} \cdots \mathbb{P}_{\alpha_N}) \pi_{\alpha_1} \cdots \pi_{\alpha_N} \\ &= \sum_{(\alpha_1, \dots, \alpha_{N-1})} [\vec{p}_1 (\mathbb{P}_{\alpha_1} \cdots \mathbb{P}_{\alpha_{N-1}}) \pi_{\alpha_1} \cdots \pi_{\alpha_{N-1}}] (\sum_{\alpha_N} \mathbb{P}_{\alpha_N} \pi_{\alpha_N}) \\ &= \vec{p}_1 \mathbb{A}^{N-1} \mathbb{A} = \vec{p}_1 \mathbb{A}^N \quad || \end{aligned}$$

This theorem could alternatively be proven as a consequence of the theorem which states that the expectation of a product of independent random variables is the product of their expectation. Then the work would be to show that the conditions of this theorem are satisfied for matrix random variables and a Bernoulli p.s.g. along with the model $\mathcal{M}$.

In the mathematical learning theory literature, the two most frequent experimental paradigms for list learning are the anticipation procedure and the R-T procedure. The anticipation procedure presents no difficulty for the framework. Thus an anticipation procedure for a list of M items could be defined in terms of a p.s.g. which selects any of the $M!$ orders of the M items with probability $1/M!$ at the start of each cycle.

The R-T procedure, however, presents slightly more difficulty for the framework. The problem is that a trial in the R-T procedure does not fit the definition on p. 54 of this chapter. Instead of the stimulus being presented for a response followed immediately by a presentation of the S-R pair, the R-T procedure groups the R-trials (presentations of S-R pairs), groups the T-trials (presentations of the S members for a response), and alternates blocks of R and T trials.

There is a fairly simple and natural way to extend the framework to handle the R-T procedure as well as several other situations to be mentioned. Suppose an event, E_ν, is defined to be any occurrence in an experiment which a theory says may affect the state of the list. Thus far, theories have been restricted to those which specify that the only events are presentations of stimuli for an anticipation trial, i.e., thus far, transitions among states of the list are permitted only upon the presentation of a stimulus. We could associate a transition matrix $\mathbb{P}_\nu$ with each event E_ν. Then, if event E_ν occurs at some time point N in the experiment, $\mathbb{P}_\nu$ would be applied to $\vec{p}_N$ to give $\vec{p}_{N+1}$, i.e.,

$$\vec{p}_{N+1} = \vec{p}_N \mathbb{P}_\nu \quad . \tag{4.15}$$

Now the R-T procedure can easily be accommodated within the framework. Associated with each type of R-trial, R_ν, is a matrix $\mathbb{P}_\nu$ and with each type of T-trial, $T_{\nu'}$, a matrix $\mathbb{P}_{\nu'}$. Thus during a T cycle the associated item matrices for T-trials are effective, and during R-trials the matrices for R-trials on items are effective. Viewed in this way, the issue of learning on test trials is whether or not the transition matrices for T-trials are diagonal (1s on the diagonal and 0s elsewhere).

Other places where this generalized notion of event might prove useful are as follows. Peterson and Peterson (1962) had subjects count backwards between an R-trial and a T-trial to study memory. If models for the Petersons' experiment were written in the framework of this chapter, one could define a counting event, E_v. Presumably the associated matrix $\mathbb{P}_v$ would tend to shift an item into a forgotten state.

A second possible use of the generalization of event in the framework would be in the optimization work of Crothers (1965). Crothers considered two types of trials (modes of presenting material to be learned) and the paper concerned finding a solution to the problem of the optimal scheduling of these trials under various constraints. To solve this problem, he associated a transition matrix with each type of event; a matrix was assumed to be effective on any trial when its associated event occurred. The members of the state space, however, were not the states of the list but were states of a particular item.

It would lead us too far astray to develop additional properties of presentation schedules within the multi-level framework. The preceding comments should indicate the way to incorporate a p.s.g. into the framework.

Before leaving the section on computing state and response probabilities for multi-level models, there is another property of some models that can simplify derivations. If the matrices in $\mathscr{P}$ commute, computations from a model are simplified (Theorem 4.3). First, we formalize the notion of a commutative model in Definition 4.7 and then state Theorem 4.3.

Definition 4.7

Suppose a list of M items and a model $\mathcal{M}$. $\mathcal{M}$ is said to be a commutative model in case, for all $\alpha, \beta = 1, 2, \ldots, M$,

$$\mathbb{P}_\alpha \cdot \mathbb{P}_\beta = \mathbb{P}_\beta \cdot \mathbb{P}_\alpha \; .$$

Examples of commutative models are the one-element P-level model and the all-or-none multi-level model. The former is commutative as a consequence of the property that each item in the list is learned independently. To see this effect on the matrices, consider the P-level model for $M = 2$. We may compute $\mathbb{P}_1 \cdot \mathbb{P}_2$ and $\mathbb{P}_2 \cdot \mathbb{P}_1$, where $\mathbb{P}_1$ and $\mathbb{P}_2$ are given by Eqs. (4.8) and (4.9) respectively. The result is

$$\mathbb{P}_1 \cdot \mathbb{P}_2 = \mathbb{P}_2 \cdot \mathbb{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 \\ c & 0 & 1-c & 0 \\ c^2 & c(1-c) & c(1-c) & (1-c)^2 \end{bmatrix} .$$

Theorem 3.3 suffices to prove that the all-or-none multi-level model is commutative. The appropriate matrix multiplication for $M = 2$ is presented in Eq. (4.4).

Commutative models make a strong prediction that the order of presenting stimuli does not matter. This is shown in Theorem 4.3.

Theorem 4.3

Suppose a list of M items, a model $\mathcal{M}$ with associated start vector $\vec{p}_1$. Suppose $\mathcal{M}$ is a commutative model and that, for $\alpha = 1, 2, \ldots, M$, k_α S_α items are presented in the first N trials, i.e.,

$$\sum_{\alpha=1}^{M} k_\alpha = N .$$

Then $\vec{p}_{N+1}$ is independent of the order of presenting the items.

Proof

From Theorem 1 and repeated use of commutativity of the members of $\mathcal{P}$ we have, for any presentation order,

$$\vec{p}_{N+1} = \vec{p}_1 \mathbb{P}_1^{k_1} \cdot \mathbb{P}_2^{k_2} \cdots \mathbb{P}_M^{k_M} \quad || \tag{4.16}$$

D. Reduction of the Cardinality of Models

Before turning to specific applications, one more aspect of the framework needs to be developed. As the size of the list, M, and the size of the item state space, L, increase, the size of $\mathcal{T}$, the state space for the list, increases as L^M. Even for a three-stage model for a twenty-item list $\mathcal{T}$ would have $3^{20} = 3{,}486{,}784{,}401$ possible states in its representation within the framework. Also, working with matrices of the order of 3.5×10^9 by 3.5×10^9 would tax the abilities of the strongest computer.

There are, however, several ways to reduce the cardinality of objects in the framework. Three of these will be developed in the next few pages. By way of preview, the first will be to drop inaccessible states, the second will be to break down a list into sublists such that no item dependencies (or mutual interactions) exist between members of separate sublists, and the third is to use the notion of lumping states in a Markov chain (cf. Burke and Rosenblatt, 1958) to effect computational simplicities in determining $\vec{p}_N$.

The first of these has already been used in this chapter for the all-or-none multi-level model with $M = 2$. States (R, P), (P, R), (R, U), and (U, R) were dropped in the matrices in Eqs. (4.2) and (4.3). This is because none of these states is obtainable from other states and further have zero probability in $\vec{p}_1$. With this in mind we state the following definition.

Definition 4.8

Suppose a list $\mathcal{I}$ of M items, and a model $\mathcal{M} = (\mathcal{T}, \mathcal{P}, \mathcal{L})$ with associated start vector $\vec{p}_1$. Define the class of null states,

η, to be the largest subset of $\mathcal{T}$ such that:

1. For all $\vec{t}_k \in \eta$

$$\vec{p}_1^{(k)} = 0 \; .$$

2. For all $\vec{t}_k \in \eta$, $\vec{s}_j \in \mathcal{T}-\eta$ and $\alpha = 1, 2, \ldots , M$,

$$\mathbb{P}_\alpha^{jk} = 0 \; .$$

For the all-or-none multi-level model,

$$\eta = \{(R, P), (P, R), (R, U), (U, R)\} \; .$$

It should be fairly obvious that if the state space of the list is taken as

$$\mathcal{T} - \eta$$

the preceding definitions and theorems are unaffected in content; hence, from now on, when a model $\mathcal{M} = (\mathcal{T}, \mathcal{P}, \mathcal{L})$ is considered, it can be assumed that η has been dropped from $\mathcal{T}$.

Dropping null states would be very important in situations where learning takes place mostly at high levels, i.e., where the model specifies that large collections of items change states at once or not at all. For example, consider the one-element R-level model for $M = 20$. $\mathcal{T}$ in the associated model would be of size 2^{20}; however, $\mathcal{T} - \eta$ would have only two members: $(U, U, \ldots , U)$ and $(R, R, \ldots , R)$.

The second method of reducing the cardinality of $\mathcal{T}$ (or $\mathcal{T} - \eta$) is illustrated by a typical P-level analysis of a paired-associate experiment. In effect, a list of size M is reduced to M lists of size one. This is possible since the transition probabilities and response probabilities for a particular item are assumed not to depend on the states of other items in the list or even the number and order of previous presentations of other items. Thus, items that depend in no way on each other in their course of learning can be analyzed as though they came from separate lists. This observation is formalized in Theorem 4.4. First, Definition 4.9 concerning the classes of item dependencies which a model might postulate is stated.

In the definition to follow, the notion of the set of items dependent on an item is developed. This idea is then used to define "level of learning," and finally, a theorem about breaking a list into independent sublists is stated.

Suppose S_α is a particular item in a list $\mathcal{S}$. By D_α is meant the set of items in $\mathcal{S}$ dependent on item S_α. An item S_β is said to be dependent on S_α in case any one of three possibilities obtain: i. response probabilities to S_β depend on the state of S_α; ii. the state of S_β can change on trials when S_α is presented; or iii. the transition probabilities for S_β, when S_α is not presented, depend on the state of S_α. These notions are formalized in Definition 4.9.

Definition 4.9

Suppose a list of M items and a model $\mathcal{M} = (\mathcal{I}, \mathcal{P}, \mathcal{L})$ with associated start vector $\vec{p}_1$. For each $\alpha = 1, 2, \ldots, M$, the sets D_α^1, D_α^2, and D_α^3 are defined as follows:

i. $D_\alpha^1 = \{S_\alpha\colon S_\beta \in \mathcal{S}$ and there exists a $A_i \in \mathcal{A}$ and $\vec{u}, \vec{v} \in \mathcal{I}$ differing only in their α^{th} position such that

$$\Pr(A_i \mid S_\beta, \vec{u}) \neq \Pr(A_i \mid S_\beta, \vec{v})\} .$$

ii. $D_\alpha^2 = \{S_\beta\colon S_\beta \in \mathscr{S}$ and there exist $\vec{u}, \vec{v} \in \mathscr{T}$ with $\vec{u}^\beta \neq \vec{v}^\beta$ such that

$$\mathbb{P}_\alpha^{\vec{u}\,\vec{v}} \neq 0\} .$$

iii. $D_\alpha^3 = \{S_\beta\colon S_\beta \in \mathscr{S}$ and there are $\vec{u}, \vec{v} \in \mathscr{T}$ differing only in their α^{th} position, $S_\gamma \in \mathscr{S}$ with $\gamma \neq \alpha$, and $\tau \in T_I$ such that

$$\Pr(\vec{t}_{N+1}^{(\beta)} = \tau \mid S_{\gamma,N}, \vec{u}_N) \neq \Pr(\vec{t}_{N+1}^{(\beta)} = \tau \mid S_{\gamma,N}, \vec{v}_N)\} .$$

Then the set of items dependent on an item S_α is defined to be

$$D_\alpha = D_\alpha^1 \cup D_\alpha^2 \cup D_\alpha^3 .$$

D_α^1 in the preceding definition is just the set of items whose response probabilities can be affected by the current state of item S_α; D_α^2 is the set of items whose states can change when S_α is presented; and D_α^3 is the set of items whose transition probabilities can be affected by the state of S_α. D_α, then, is the set of items dependent on S_α in any of these senses.

In the next section, Definition 4.9 will be used to define a dependency relation on $\mathscr{S}$. This relation will be extended to an equivalence relation in order to define level of learning in terms of the partition of $\mathscr{S}$ induced by the extended dependency relation.

Let us define a binary relation, D, on $\mathscr{S}$ in terms of the D_α's. We say S_β depends on S_α, written $S_\beta D S_\alpha$ in case $S_\beta \in D_\alpha$ for $\alpha, \beta = 1, 2, \ldots, M$. Anticipating the development to follow, it would be a desirable property if $\{D_\alpha\colon \alpha = 1, \ldots M\}$ forms a partition of $\mathscr{S}$, i.e., if the D_α are mutually exclusive and $\bigcup_\alpha D_\alpha = \mathscr{S}$. Put another way, it would be desirable if the relationship of item dependency was an equivalence relation, i.e., reflexive, symmetric, and transitive. In this case, as a later theorem will show, the list could conveniently be broken into sublists, and each subject learning the list would provide one set of data for each sublist.

Examples are the one-element P-level model where $D_\alpha = \{S_\alpha\}$ and the all-or-none multi-level model where the equivalence classes are the groups of M related items.

It would be unduly restrictive to require $\{D_\alpha\}$ to be a partition of $\mathscr{S}$. Consider the mixed model (Atkinson and Estes, 1963) for the following list:

stimulus	response
ab	1
bc	2
cd	3

For this case $D_{ab} = \{ab, bc\}$, $D_{bc} = \{ab, bc, cd\}$ and $D_{cd} = \{bc, cd\}$--certainly not a partition of $\mathscr{S}$. These results come from the fact that the conditioning axioms for the mixed model require, for all stimuli x, $D_x^2 = \{x\}$ and $D_x^3 = \{x\}$; however, D_x^1 is the set of all stimuli which share components with x. The dependency relation is reflexive and symmetric for the mixed model, but not necessarily transitive. (Parenthetically, one could test transitivity for such a list by manipulating presentation sequences and observing whether preceding presentations of *ab* affect response probabilities to *cd*. Although this experimental question is of interest to the author, it will not be pursued in this paper.)

Since it is too restrictive to assume that the dependency relation D is an equivalence relation, we define an extension of D to an equivalence relation and base an operation of breaking a list into sublists based on this extended equivalence relation.

Definition 4.10

Suppose a list $\mathscr{S}$ of M items, a model $\mathscr{M}$ with associated start vector $\vec{p}_1$, and the dependency relation D induced by $\mathscr{M}$. Define D^*, the levels extension of D, to be the minimal equivalence relation containing D, i.e., $D^* - D$ has the fewest members.

Strictly speaking, the preceding requires a result from set theory to be a proper definition. Clearly, $\mathscr{S} \times \mathscr{S}$ is an equivalence relation containing D. Hence, setting D^* equal to the intersection of all such equivalence relations containing D, which is easily shown to yield an equivalence relation, suffices to establish the existence of D^*. Since $\mathscr{S}$ is assumed finite, D^* can be easily obtained by construction. One simply adds to D all pairs from $\mathscr{S} \times \mathscr{S}$ necessary to satisfy symmetry, transitivity, and the reflexive property. Denote by $\mathscr{S}^*$ the partition of $\mathscr{S}$ induced by D^*.

To illustrate the preceding process, consider the mixed model for the following two lists:

LIST 1

stimulus	response
ab	1
bc	2
cd	3
ef	4

LIST 2

stimulus	response
ab	1
bc	2
cd	3
de	4
ef	5

For List 1 we have

$$D' = \{(ab, ab), (bc, bc), (cd, cd), (ef, ef), (bc, ab), (ab, bc), (cd, bc), (bc, cd)\};$$

$$D'^{*} = \{ab, bc, cd\} \times \{ab, bc, cd\} \cup \{(ef, ef)\},$$

in other words the pairs (ab, cd) and (cd, ab) have been added to D' to satisfy transitivity; and finally

$$\mathcal{D}'^{*} = \{\{ab, bc, cd\}, \{ef\}\}.$$

However, for list two,

$$D'' = D' \cup \{(de, de), (ef, de), (cd, de), (de, ef), (de, cd)\};$$

$$D''^{*} = \{ab, bc, cd, de, ef\} \times \{ab, bc, cd, de, ef\};$$

and

$$\mathcal{D}''^{*} = \{\{ab, bc, cd, de, ef\}\}.$$

In other words the addition of item de to List 1 is sufficient to tie all the stimuli in the list together in the sense of Definition 4.10.

We are finally in the position to offer a possible definition of the notion of "level of learning" used extensively in Chapters 2 and 3. By a level of learning is meant a partition of $\mathcal{S}$, i.e., a collection of subsets of $\mathcal{S}$ which are mutually disjoint and exhaustive. By the highest level of learning for a list and model $\mathcal{M}$ is meant the finest partition (one with the most equivalence classes) of $\mathcal{S}$ for which items in different subsets are mutually independent, that is, if $S_\alpha \in A \subset \mathcal{S}$ and $S_\beta \in B \subset \mathcal{S}$, and if $A \neq B$, then not $S_\alpha D S_\beta$ and not $S_\beta D S_\alpha$. It will turn out that $\mathcal{D}^{*}$ is the appropriate partition.

In the theorem to follow we will base a method of breaking a list into sublists corresponding to the equivalence classes of $\mathcal{D}^{*}$. At the same time the list is broken into sublists the model $\mathcal{M}$ can be broken into corresponding submodels. The procedure is analogous to breaking one long string of errors and successes into a group of short sequences, one for each item, as is done in P-level analyses. The important property, captured in the theorem, is that presentations of items outside of a fixed cell in $\mathcal{D}^*$ act as "dead trials" relative to changes of state probabilities and response probabilities of members of the equivalence class.

Unfortunately, the theorem, although fairly intuitive, becomes very cumbersome from a notational standpoint. Therefore, some of the notation used in the theorem will be defined as follows.

Let $\mathcal{S}_N$ be the set of all possible presentation sequences through trial N. Then

$$\mathcal{S}_N = \underbrace{\mathcal{S} x \ldots x \mathcal{S}}_{N\text{-times}} .$$

For each $B_i \in \mathcal{D}^*$ define $\mathcal{S}_N^i$ to be the set of all sequences of presentations from B_i for the first N trials. Then

$$\mathcal{S}_N^i = \underbrace{B_i x \ldots x B_i}_{N\text{-times}} .$$

Each sequence $s_N \in \mathcal{S}_N$ can be decomposed into subsequences, such that a particular subsequence, s_N^i, represents all presentations of members of a particular B_i. Thus, for each $s_N \in \mathcal{S}_N$, s_N^i is a subsequence of length p, $0 \leq p \leq N$, consisting only of the members of B_i listed in s_N. Finally, for each $\vec{t} \in \mathcal{T}$, define $\vec{t}(i)$ to be the set of all $\vec{u} \in \mathcal{T}$ which agree with $\vec{t}$ in coordinate positions corresponding to the states of items in B_i, i.e.,

$$\vec{t}(i) = \{\vec{u}: \vec{u} \in \mathcal{T} \text{ and } \vec{u}^{\alpha} = \vec{t}^{\alpha} \text{ for all } \alpha \text{ such that } S_\alpha \in B_i\} .$$

It should be clear that for each $B_i \in \mathcal{S}^*$ $\{\vec{t}(i): \vec{t} \in \mathcal{T}\}$ forms a partition of $\mathcal{T}$.

With these notational devices in mind, we are prepared to state the major theorem of this section. Theorem 4.4 asserts that the response probabilities to an item in a particular equivalence class, $B_i \in \mathcal{S}^*$, depend only on the order and number of preceding items in that equivalence class.

Theorem 4.4

Suppose $\mathcal{S}$ is a set of M items with model $\mathcal{M} = (\mathcal{T}, \mathcal{P}, \mathcal{L})$ and associated start vector $\vec{p}_1$. Let $\mathcal{S}^* = \{B_1, B_2, \dots , B_\nu\}$ be the levels partition of $\mathcal{S}$ induced by the equivalence relation D*. Then, for all $B_i \in \mathcal{S}^*$, $N = 1, 2, \dots$, $s_N \in \mathcal{S}_N$, and responses $A_j \in \mathcal{A}$,

$$\Pr(A_{j,N+1} | S_{\alpha,N+1}, s_N) = \Pr(A_{j,p+1} | S_{\alpha,p+1}, s_N^i) ,$$

for $i = 1, 2, \dots , \nu$ and $S_\alpha \in B_i$.

Proof

$$\Pr(A_{j,N+1} | S_{\alpha,N+1}, s_N) = \sum_{\vec{t} \in \mathcal{T}} \Pr(A_{j,N+1} | S_{\alpha,N+1}, \vec{t}_{N+1}) \Pr(\vec{t}_{N+1} | s_N) .$$

By assumption of $S_\alpha \in B_i$, response probabilities to S_α depend only on the states of items in B_i, hence we have

$$\Pr(A_{j,N+1} | S_{\alpha,N+1}, s_N) = \sum_{\vec{t}(i)} \Pr(A_{j,N+1} | S_{\alpha,N+1}, \vec{t}_{N+1}(i)) \Pr(\vec{t}_{N+1}(i) | s_N)$$

where the sum is over the equivalence classes $\vec{t}(i)$ induced by B_i, i.e., the subsets $\mathcal{J}$ corresponding to each sequence of states of items in B_i. The next step involves noting that for all $s_N \in \mathcal{S}_N$,

$$\Pr(\vec{t}_{N+1}(i)|s_N) = \Pr(t_{p+1}(i)|s_N^i) \ .$$

This is established by summing over members of the set $\vec{t}_{N+1}(i)$ as follows:

$$\begin{aligned}
\Pr(\vec{t}_{N+1}(i)|s_N) &= \sum_{\vec{u}\in\vec{t}(i)} \Pr(\vec{u}_N|s_N) \\
&= \sum_{\vec{u}\in\vec{t}(i)} [\vec{p}_1 \prod_{k\in s_N} \mathbb{P}_k]^{(\vec{u})} \\
&= \sum_{\vec{u}'\in\vec{t}(i)} [\vec{p}_1 \prod_{k'\in s_N^i} \mathbb{P}_{k'}]^{(\vec{u}')} \\
&= \Pr(\vec{t}_{p+1}(i)|s_N^i) \ .
\end{aligned}$$

The last two steps follow, since the states of items in B_i can change only on presentations of items in B_i, and, further, the transition probabilities for items in B_i do not depend on the states of items not in B_i. The reader not convinced of this should note that, for all $S_\beta \notin B_i$, and any $S_\alpha \in B_i$, and $\vec{u}, \vec{v} \in \mathcal{J}$ with $\vec{u}^{\alpha} \neq \vec{v}^{\alpha}$

$$\mathbb{P}_\beta^{\vec{u}\vec{v}} = 0 \ .$$

Putting these results together yields

$$\Pr(A_{i,N+1}|S_{\alpha,N+1}, s_N) = \Pr(A_{i,p+1}|S_{\alpha,p+1}, s_N^i) \ \|$$

The gist of this theorem is that presentation of items not in a cell $B_i \in \mathcal{B}^*$ do not affect the states or response probabilities of items in B_i. Consequently, each cell of $\mathcal{B}^*$ can be studied as an independent sublist of $\mathscr{S}$. The associated model $\mathscr{M}_i$ for the sublist corresponding to $B_i \in \mathscr{B}^*$ will have state space $\mathscr{T}_i$, where

$$\mathscr{T}_i = \{t(i): t \in \mathscr{T}\} .$$

If B_i has M_i members, then $\mathscr{T}_i$ will have L^{M_i} members. Finally, the transition probabilities,

$$\mathbb{P}_{\alpha}^{\vec{t}(i)\vec{t}(j)} ,$$

are determined from $\mathscr{M}$ since

$$\mathbb{P}_{\alpha}^{\vec{u}\vec{v}}$$

is the same for all pairs $\vec{u} \in \mathscr{T}(i)$, $\vec{v} \in \mathscr{T}(j)$.

Thus far we have discussed how inaccessible states in $\mathscr{T}$ (namely, the η states) can be eliminated. Also we have considered how, in certain cases, a list of M items can be partitioned into shorter sublists with a consequent reduction in the number of states needed to characterize the learning of each sublist. A third possibility for reducing the size of $\mathscr{T}$ (or $\mathscr{T} - \eta$ or $\mathscr{T}_i$ for $B_i \in \mathscr{B}^*$) is to reduce the number of states in the item state space T_I. This operation would reduce L and hence L^M. In this section we consider briefly the notion of lumping (combining) certain of the states in $\mathscr{T}$. "Lumping" is a technical topic in the Markov chain literature (cf. Burke and Rosenblatt, 1958; Kemeny and Snell, 1960, pp. 132-140), and its use within the framework is a highly model-specific question.

The basic idea of lumping (or combining) the states of a Markov chain is as follows. Suppose M is a Markov chain with state space

$X = \{x_1, x_2, \ldots, x_n\}$. Let $Y = \{y_1, y_2, \ldots, y_m\}$ be a partition of X, i.e., Y consists of pair-wise disjoint subsets of X whose union is X. If the state space Y forms a Markov chain, then Y is said to be a valid lumping of the states in X. A sufficient condition for the lumped process Y to represent the state space for a Markov chain is as follows (Burke and Rosenblatt, 1958): Y is a valid lumping of X in case, for each $y_i, y_j \in Y$

$$Pr(x-y_j) = Pr(x'-y_j)$$

for all $x, x' \in y_i$, where $Pr(x-y_j)$ is the probability of transition from state $x \in X$ to the class of states $y_j \subset X$. It should be noted that if this condition is satisfied $Pr(Y_i-Y_j)$ is well defined--otherwise it is not.

Depending on the model, this condition can be used to lump the states in each $P \in \mathcal{P}$ in such a way as to reduce the size of $\mathcal{S}$. In most cases this method of reducing $\mathcal{S}$ depends on the particular model; however, there is one case for which a somewhat general condition permitting lumping of the states in $\mathcal{S}$ can be established. In the case where all M items in a list are similar in a sense to be established in Definition 4.11, it is possible to lump the states of $\mathcal{S}$ if each item is equally likely to appear on any trial. This condition is established in Theorem 4.5. First we formalize the notion of a "symmetric model" which plays a role in this theorem.

Definition 4.11

Suppose a list $\mathcal{I}$ of M items and a model $\mathcal{M} = (\mathcal{S}, \mathcal{P}, \mathcal{L})$. $\mathcal{M}$ is said to be symmetric in case $\mathcal{P}$ is left unchanged by permuting the order of stimuli listed in the state space $\mathcal{T}$.

A symmetric model has the property that all items are treated alike by the model. By this is meant that the same set of matrices, $\mathcal{P}$, would be obtained for any ordering of the items in the definition of the state space of the list, i.e., there is a matrix in $\mathcal{P}$ associated with the first listed item in the state space, one associated with the second, ... ; and further, these matrices do not depend on which item is listed first, second, ... As an example, consider the one-element P-level model for a two-item list. Let S and S' denote the two items. Eqs. (4.8) and (4.9) give the two members of $\mathcal{P}$ for the S'S order of listing the states of the items. It should be clear that if the items were listed as SS' in the state space, the same two matrices would be obtained, except Eq. (4.8) would apply on S trials and Eq. (4.9) on S' trials instead of the reverse set-up for the S'S order of listing states of items in $\mathcal{T}$.

All of the models considered in this chapter are symmetric models in the sense of Definition 4.11. One type of consideration that would tend to make a model asymmetric would be items of unequal difficulty. To illustrate, consider the aforementioned one-element P-level model for the list $\mathcal{L} = \{S,S'\}$; however, suppose S is an easier item to learn than S'. Accordingly, let c be greater than c'. Suppose the state of the list is listed in the SS' order, then the resulting matrices are as follows:

$$(4.17) \qquad P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ c & 0 & 1-c & 0 \\ 0 & c & 0 & 1-c \end{bmatrix}$$

and

$$(4.18) \qquad \mathbb{P}' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ c' & 1-c' & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & c' & 1-c' \end{bmatrix};$$

However, if S' is listed before S in $\mathcal{J}$, the resulting matrices are

$$(4.19) \qquad \mathbb{P}' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ c' & 0 & 1-c' & 0 \\ 0 & c' & 0 & 1-c' \end{bmatrix}$$

and

$$(4.20) \qquad \mathbb{P} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & c & 1-c \end{bmatrix}.$$

Since these two sets of matrices (Eqs. (4.17) and (4.18) vs. Eqs. (4.19) and (4.20)) are not the same, the model is not symmetric. One important thing to note about requiring a model to be symmetric is that it places no restriction on what types of item dependencies are possible, e.g., the all-or-none multi-level model is a symmetric model.

If a model is symmetric and is tested by a Bernoulli presentation schedule with $\pi_i = 1/M$ for $i = 1, 2, \ldots, M$ (see Definition 4.6), it is possible to lump the states of the average matrix (see Theorem 4.2),

$$(4.21) \qquad A_{\frac{1}{M}} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{P}_i .$$

The lumping permitted by these conditions produces a partition of $\mathcal{T}$, $\mathcal{E}$, defined as follows. Suppose a list, $\mathcal{S}$, of M items with item state space $T_I = \{\tau_1, \tau_2, \dots \tau_L\}$. Then $\mathcal{E}$, the counting partition of $\mathcal{T}$, is defined as follows:

(4.22) $\mathcal{E} = \{\vec{e} = (e_1, e_2, \dots, e_L)$: e_i is a number between 0 and M representing the number of items in state τ_i for $i = 1, 2, \dots, L$; and $\sum_{i=1}^{L} e_i = M\}$.

Actually $\mathcal{E}$ itself is not a partition of $\mathcal{T}$, but corresponding to each $\vec{e} \in \mathcal{E}$, there is a subset e of $\mathcal{T}$ whose vectors have e_i items in state τ_i ($i = 1, 2, \dots, L$). It is these subsets, e, which form a partition of $\mathcal{T}$ and correspond to the states of the lumped process presented in Theorem 4.5.

Theorem 4.5

Suppose $\mathcal{M} = (\mathcal{T}, \mathcal{P}, \mathcal{L})$ is a symmetric model for a list of M items. Suppose that items are presented with a Bernoulli presentation schedule with $\pi_i = 1/M$ for $i = 1, 2, \dots, M$. Then the Markov chain with state space $\mathcal{T}$ and stochastic matrix given by Eq. (4.21) lumps to a Markov chain with state space $\mathcal{E}$.

Proof

The proof proceeds by using the Burke-Rosenblatt criterion discussed on p. 82 of this chapter. Let e and e' be any two sets in $\mathcal{E}$. Then, for all $\vec{t}_i, \vec{t}_j \in e$, we have

$$\Pr(\vec{t}_i - e') = \Pr(\vec{t}_j - e') \quad (4.23)$$

Eq. (4.23) comes from the fact that $\mathcal{M}$ is symmetric. That is

$$\begin{aligned}
\Pr(\vec{t}_i\text{-}e') &= \sum_{k=1}^{M} \Pr(\vec{t}_i\text{-}e' \mid S_k)\Pr(S_k \mid \vec{t}_i) \\
&= \frac{1}{M}\sum_{k=1}^{M} \Pr(\vec{t}_i\text{-}e' \mid S_k) \\
&= \frac{1}{M}\sum_{k=1}^{M} \Pr(\vec{t}^{\,*}\text{-}e' \mid S_k) \;,
\end{aligned}$$

where $\vec{t}^{\,*}$ is the vector corresponding to $\vec{t}_i$ when the order of items in the state space $\mathcal{T}$ is permuted to list items in state τ_1 first, in τ_2 second, ... , and items in τ_L last. Since the model is symmetric, the resulting set of matrices $\mathcal{P}*$ is the same as $\mathcal{P}$ and hence the above equation holds. A similar argument implies

$$\Pr(\vec{t}_j\text{-}e') = \frac{1}{M}\sum_{k=1}^{M} \Pr(\vec{t}^{\,*}\text{-}e' \mid S_k) \;,$$

and hence Eq. (4.23) follows. This result establishes that the Burke-Rosenblatt criterion holds; and hence, the lumping of $\mathcal{T}$ to $\mathcal{E}$ is valid. ||

Next, we indicate how this theorem, along with Theorem 4.2, might be used in the analysis of a particular symmetric model. Theorem 4.5 and Theorem 4.2 are used in several places in Chapter 5 to derive particular models from a general model in the sense of Definition 4.3 (see pp. 112-117 for Restle's strategy-selection theory). Consider the one-element P-level model for $M = 2$. The two matrices, $\mathbb{P}_1$ and $\mathbb{P}_2$, are given by Eqs. (4.8) and (4.9). Suppose a Bernoulli presentation schedule with $\pi_1 = \pi_2 = \frac{1}{2}$ (see Definition 4.6), then $A_{\frac{1}{2}}$, for Theorem 4.2, is as follows:

$$(4.24)\qquad \mathbb{A}_{\frac{1}{2}} = \begin{array}{c} \\ (L,L) \\ (L,U) \\ (U,L) \\ (U,U) \end{array}\!\!\begin{array}{c} \begin{array}{cccc} (L,L) & (L,U) & (U,L) & (U,U) \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ \frac{1}{2}c & 1-\frac{1}{2}c & 0 & 0 \\ \frac{1}{2}c & 0 & 1-\frac{1}{2}c & 0 \\ 0 & \frac{1}{2}c & \frac{1}{2}c & 1-c \end{array}\right] \end{array}.$$

Since the model is symmetric (see p. 82), we can use Theorem 4.5 to lump the state space $\mathcal{J} = \{(L,L), (L,U), (U,L), (U,U)\}$. The counting partition, $\mathcal{E}$, is given by

$$(4.25)\qquad \mathcal{E} = \{\{(L,L)\}, \{(L,U), (U,L)\}, \{(U,U)\}\}$$

(denote these three sets by e_2, e_1, and e_0, respectively). Using Theorem 4.5, we obtain the following stochastic matrix for the lumped process with state space $\mathcal{E}$:

$$(4.26)\qquad \mathbb{A}^{L}_{\frac{1}{2}} = \begin{array}{c} \\ e_2 \\ e_1 \\ e_0 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} e_2 & e_1 & e_0 \end{array} \\ \left[\begin{array}{ccc} 1 & 0 & 0 \\ \frac{1}{2}c & 1-\frac{1}{2}c & 0 \\ 0 & c & 1-c \end{array}\right] \end{array}.$$

This derived matrix can be used as a Markov model to describe the error-success process on the pair of items $\{S_1,S_2\}$ under a Bernoulli presentation schedule with $\pi_1 = \pi_2 = 1/2$. A model for error-success sequences is conventionally displayed as a transition matrix among theoretical states along with a column matrix of the probability of a correct response given a particular state (cf. Atkinson, Bower, and Crothers, 1965, pp. 89, 253, and 305). Accordingly, the model for error-success sequences on $\{S_1, S_2\}$ derived from the one-element P-level model is displayed in Eq. (4.27) as follows:

$$(4.27)\quad \begin{array}{c} \\ e_{2,n} \\ e_{1,n} \\ e_{0,n} \end{array} \begin{array}{c} \begin{array}{ccc} e_{2,n+1} & e_{1,n+1} & e_{0,n+1} \end{array} \\ \left[\begin{array}{ccc} 1 & 0 & 0 \\ \frac{1}{2}c & 1-\frac{1}{2}c & 0 \\ 0 & c & 1-c \end{array}\right] \end{array} \qquad \begin{array}{c} \Pr(\text{correct}|\text{row state}) \\ \left[\begin{array}{c} 1 \\ \frac{1}{2}(1+g) \\ g \end{array}\right] \end{array},$$

where g is the probability of a correct response for a presented item in state U. It is of some interest to note that the model in Eq. (4.27) is formally identical to the two-element pattern model axiomatized by Suppes and Atkinson (1960, pp. 14-17). The equivalence comes from interpreting the stimuli S_1 and S_2 as the two patterns. The Bernoulli schedule employed guarantees that each stimulus (pattern) is sampled on each trial with probability 1/2. It is interesting to note that Suppes, et al. (1962) lumped the four-state matrix for the two-element model to one equivalent to Eq. (4.26). The preceding observations suggest that, if the particular stimulus giving rise to an error or success on trial n is suppressed in the level of analysis corresponding to the analysis of the error-success process on M items for a Bernoulli presentation schedule, then the resulting model bears a resemblance to an M-element pattern model. However, in the case of more complex multi-level models than the one-element P-level model, it is possible for "patterns" (stimuli) to change their states when not "sampled" (presented).

One final comment about the model represented by Eq. (4.27) is needed. Transitions from state e_1 to e_2 have different probabilities following an error than following a success. This is because an error in

e_1 implies the unlearned item has been sampled; whereas, a success in e_1 does not determine the state of the presented stimulus. This feature is shared by the two-element pattern model of Suppes, et al. (1962). Although analyses of the two-element model can be found in the literature (cf. Bower and Theios, 1963), in general, there may be more than two items in the list. When $M > 2$, analysis of the resulting model (obtained by Theorems 4.2 and 4.5) is best done by computer.

Bernbach (1966) has proposed a computerizable scheme for analyzing Markov models. To use Bernbach's scheme, it is necessary to expand each state into an error and a success state. When this expansion is accomplished, the differential probability of learning after an error or success is embodied in the matrix. To illustrate, Eq. (4.27) can be so expanded; the result is as follows:

$$(4.28) \quad \begin{array}{c|ccccc|c} & e_2 & e_1^S & e_1^E & e_0^S & e_0^E & \Pr(\text{correct} \mid \text{row state}) \\ \hline e_2 & 1 & 0 & 0 & 0 & 0 & 1 \\ e_1^S & A & (1-A)g' & (1-A)(1-g') & 0 & 0 & 1 \\ e_1^E & c & (1-c)g' & (1-c)(1-g') & 0 & 0 & 0 \\ e_0^S & 0 & cg' & c(1-g') & (1-c)g & (1-c)(1-g) & 1 \\ e_0^E & 0 & cg' & c(1-g') & (1-c)g & (1-c)(1-g) & 0 \end{array},$$

where $g' = \frac{1}{2}(1+g)$ and

$$(4.29) \quad A = \Pr(e_{2,n+1} \mid e_{1,n}, x_n = 0)$$

$$= \frac{cg}{1+g} .$$

Bernbach's scheme could be directly applied to Eq. (4.28) to generate the statistics for the error-success on the item pair $\{S_1,S_2\}$.

In the next chapter we present some detailed analyses of several models, using the theorems and methods developed for the framework presented in this chapter. It should be emphasized that the tractability of a model within the framework depends on the construction of clever experiments designed to reduce the state space $\mathcal{T}$, and consequently the matrices in $\mathcal{P}$, to manageable proportions.

CHAPTER 5

APPLICATIONS OF THE FRAMEWORK

In this chapter several applications of the preceding framework will be presented. It is hoped that these examples will illustrate the flexibility of the approach to the problem of levels of learning presented in Chapter 4. The framework is applied to the mixed model, the all-or-none multi-level model, and Restle's strategy-selection theory.

The Mixed Model

Atkinson and Estes (1963) develop the mixed model for the learning of the following miniature list:

Stimulus	Response
ab	1
bc	2

The assumptions are that each pattern is in a state U or state L, where L is an absorbing state, and items are presumed to start in U. Responses to an item in U are governed by the stimulus components in the sense that if a pattern is in L, then its components are assumed to be connected to the response associated with that pattern. Thus, if ab is in U and bc is in L, the probability of response 1 to ac is $1/2 \times 1/2 + 1/2 \times 0 = 1/4$, where with probability $1/2$ the S uses component a, which is unconnected to either response 1 or 2, and with probability $1/2$ he uses c, which is, by assumption, connected to response 2. The one-element P-level model is assumed to govern the learning of patterns, hence dependencies among items are produced only by the response rules which are based on the conditioning of common components.

The authors assume a Bernoulli presentation schedule (see p. 66, Chapter 4) with $\pi = 1/2$. They derive a sort of "average" matrix of transition probabilities among the states of the list, (U, U), (U, L), (L, U), (L, L). It is as follows:

$$
\begin{array}{c|cccc|}
 & (L,L) & (L,U) & (U,L) & (U,U) \\
\hline
(L,L) & 1 & 0 & 0 & 0 \\
(L,U) & \frac{1}{2}c & 1-\frac{1}{2}c & 0 & 0 \\
(U,L) & \frac{1}{2}c & 0 & 1-\frac{1}{2}c & 0 \\
(U,U) & 0 & \frac{1}{2}c & \frac{1}{2}c & 1-c \\
\hline
\end{array} \; . \tag{5.1}
$$

This matrix is raised to the nth power to get state probabilities, where n indexes presentations of either stimulus.

For the record, the response probabilities given the item presented and the state of the list are as follows:

(5.2)

$Pr(A_1)$	Stimulus presented	State of list
1	ab	LL
1	ab	LU
1/4	ab	UL
1/2	ab	UU
0	bc	LL
3/4	bc	LU
0	bc	UL
1/2	bc	UU

Now let us analyze this example within the framework of the preceding chapter[5]. The list, $\mathcal{S}$, consists of two members, ab and bc. The response set $\mathcal{A}$ consists of 1 and 2. The item state space $T_I = \{U, L\}$. $\mathcal{T} = T \times T$. $\mathcal{P}$ consists of the following two matrices, where ab $= S_1$ and bc $= S_2$,

$$(5.3) \qquad \mathbb{P}_1 = \begin{array}{c} \\ (L,L) \\ (L,U) \\ (U,L) \\ (U,U) \end{array} \begin{array}{c} \begin{array}{cccc} (L,L) & (L,U) & (U,L) & (U,U) \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ c & 0 & 1-c & 0 \\ 0 & c & 0 & 1-c \end{array}\right] \end{array}$$

and

$$(5.4) \qquad \mathbb{P}_2 = \begin{array}{c} \\ (L,L) \\ (L,U) \\ (U,L) \\ (U,U) \end{array} \begin{array}{c} \begin{array}{cccc} (L,L) & (L,U) & (U,L) & (U,U) \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & c & 1-c \end{array}\right] \end{array} .$$

The response rule $\mathcal{L}$ is presented in Eq. (5.2). $\vec{p}_1 = (0,0,0,1)$, and $\mathcal{D}^* = \{\{ab, bc\}\}$ (see p. 76, Chapter 4).

The model $\mathcal{M} = (\mathcal{T}, \mathcal{P}, \mathcal{L})$ is a commutative model since

$$(5.5) \qquad \mathbb{P}_1 \cdot \mathbb{P}_2 = \mathbb{P}_2 \cdot \mathbb{P}_1 = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 \\ c & 0 & 1-c & 0 \\ c^2 & c(1-c) & c(1-c) & (1-c)^2 \end{array}\right] .$$

---

[5] Portions of this analysis are reported in Batchelder, Bjork, and Yellott (1966, Ch. 8, problem 8.G.2).

Next we apply the theorems of Chapter 4 to analyze the mixed model in terms of the framework. First, suppose k_1 S_1 and k_2 S_2 presentations in any order for the first $k_1 + k_2 = N$ trials. Then, according to Theorem 4.3 for commutative models, we have

$$\vec{p}_{N+1} = \vec{p}_1 \cdot P_1^{k_1} \cdot P_2^{k_2}$$

$$= (0,0,0,1) \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1-(1-c)^{k_1} & 0 & (1-c)^{k_1} & 0 \\ 0 & 1-(1-c)^{k_1} & & (1-c)^{k_1} \end{bmatrix}$$

$$\cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1-(1-c)^{k_2} & (1-c)^{k_2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1-(1-c)^{k_2} & (1-c)^{k_2} \end{bmatrix}$$

$$= \Big([1-(1-c)]^{k_1}[1-(1-c)^{k_2}],\ [1-(1-c)^{k_1}](1-c)^{k_2},$$

$$[1-(1-c)^{k_2}](1-c)^{k_1},\ (1-c)^{k_1+k_2}\Big) \quad .$$

Response probabilities can be easily determined using Eq. (4.11) and the response rules of Eq. (5.2). Constructing an experiment by varying the presentation order of the S_1 and S_2 stimuli would provide a strong test for the mixed model.

Next we consider a Bernoulli presentation schedule with $\pi = Pr(S_1)$. To use Theorem 4.2, we first compute $\mathbf{A}_\pi$. The result is

$$(5.7) \qquad A_\pi = \begin{bmatrix} 1 & 0 & 0 & 0 \\ (1-\pi)c & 1-(1-\pi)c & 0 & 0 \\ \pi c & 0 & 1-\pi c & 0 \\ 0 & \pi c & (1-\pi)c & (1-c) \end{bmatrix} .$$

Now we use Theorem 4.2 to determine $E_\pi(\vec{p}_{N+1})$. The result is

$$E_\pi(\vec{p}_{N+1}) = \vec{p}_1 A_\pi^N$$

$$(5.8) \qquad = (0,0,0,1)\cdot\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1-[1-(1-\pi)c]^N & [1-(1-\pi)c]^N & 0 & 0 \\ 1-(1-\pi c)^N & 0 & (1-\pi c)^N & 0 \\ \begin{matrix} 1+(1-c)^N \\ -[1-(1-\pi)c]^N \\ -[1-\pi c]^N \end{matrix} & \begin{matrix} [1-(1-\pi)c]^N \\ -(1-c)^N \\ \end{matrix} & \begin{matrix} (1-\pi c)^N \\ -(1-c)^N \\ \end{matrix} & \begin{matrix} (1-c)^N \\ \\ \end{matrix} \end{bmatrix}$$

$$= (1+(1-c)^N - [1-(1-\pi)c]^N - [1-\pi c]^N,\ [1-(1-\pi)c]^N - (1-c)^N,$$

$$(1-\pi c)^N - (1-c)^N,\ (1-c)^N) .$$

Response probabilities are easily determined using Eq. (5.2).

The tie-in to Atkinson and Estes' analysis comes from noting that for $\pi = 1/2$ Eq. (5.8) reduces to the matrix in Eq. (5.1); i.e., $A_{\frac{1}{2}}$ is Eq. (5.1). Theorem 4.2 provides a justification for considering powers of this matrix to get state probabilities under the $\pi = 1/2$ Bernoulli presentation schedule. Atkinson and Estes' choice of a single matrix determines the unit of analysis for the miniature list to be the error-success process on the pair. From this they are able to show that performance prior to the last error on the pair falls in the interval

(1/2, 5/8). This is because the stimulus not responsible for the last error can either be learned or not prior to the last error on its partner. Then the response rules specify the end points of the above interval.

The method of analysis proposed in this paper has the advantage that the items *ab* and *bc* can be analyzed separately. One consequence is that the probability of an error response prior to the last error on a particular item will be an *increasing function* of the trial index; i.e., if n indexes the presentations of *ab*, $\Pr(x_n = 1 | L > n)$ is an increasing function from a value of $1/2$ to $3/4$. This is because the mixed model assumes that learning the patterns takes place independently so, as n increases, the probability that *bc* is learned increases with consequent negative transfer to *ab*. This result comes from the analysis in this paper by noting that, under the Bernoulli presentation schedule with $\pi = \Pr(\underline{ab})$,

$$\Pr(\vec{t}_N = (U,L) | \vec{t}_N^{\,1} = u) = \frac{\Pr(\vec{t}_N=(U,L))}{\Pr(\vec{t}_N=(U,U))+\Pr(\vec{t}_N=(U,L))} \tag{5.9}$$

$$= \frac{(1-\pi e)^{N-1} - (1-e)^{N-1}}{(1-\pi e)^{N-1}} ,$$

where the appropriate probabilities from $\vec{p}_1 A^{n-1}$ in Eq. (5.8) are inserted into Eq. (5.9). Eq. (5.9) tends to 1 as N increases. Since $L_{\underline{ab}} > N$ (last error on $\underline{ab} > N$) implies

$$\vec{t}_N \in \{(U,U), (U,L)\} ,$$

the assertion of the preceding paragraph follows.

Finally, the generalization from $\pi = 1/2$ to $\pi \in (0,1)$ permits an additional powerful test of the model. Depending on the value of e, the probability of a correct response to item bc could even decrease for large values of the parameter π. Using the response probabilities from Eq. (5.2) and Eq. (5.8) yields

$$\begin{aligned} Pr(A_{2,N+1}|S_{2,N+1}) &= 1\cdot(1-[1-(1-\pi)e]^N) + \tfrac{1}{4}([1-(1-\pi)e]^N - (1-e)^N) \\ &\qquad + \tfrac{1}{2}(1-e)^N \\ &= 1 - \tfrac{3}{4}\,[1-(1-\pi)e]^N + \tfrac{1}{4}\,(1-e)^N . \end{aligned}$$

The preceding remark can be illustrated by plotting $Pr(A_{2,N}|S_{2,N})$ for $\pi = .95$ and $e = .5$. This is shown in Fig. 5.1.

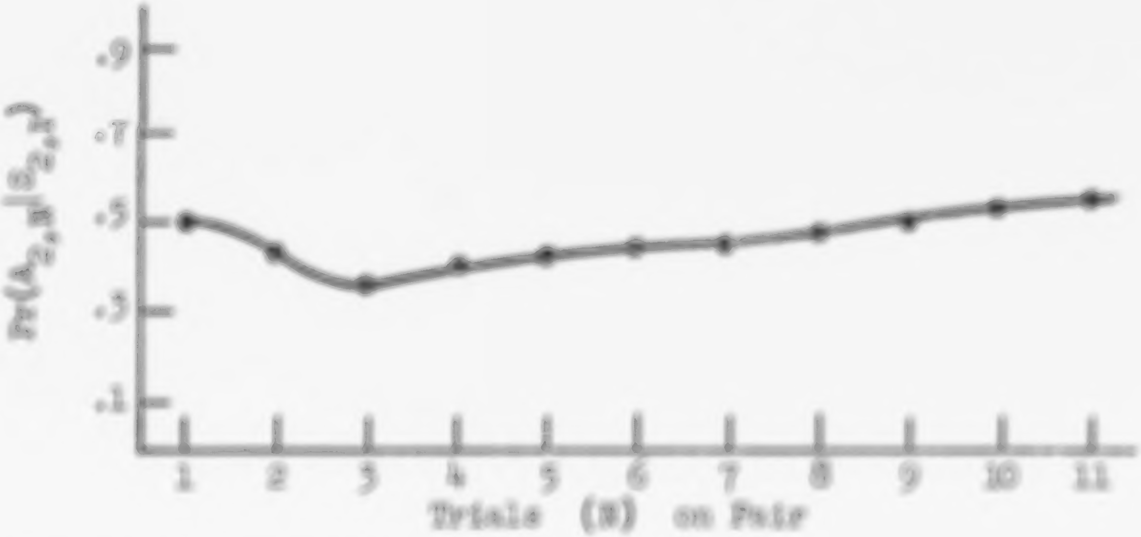


Fig. 5.1. Probability of correct to S_2 for mixed model, $\pi = .95$, $e = .50$.

This concludes the section on the applicability of the framework to the mixed model. Of course the framework could be used to gt results for other miniature lists. To recapitulate the advantages of aplying the framework to the analysis of the mixed model, we first note that properties of the model such as commutativity are utilized by tle framework (Eq. (5.6)). Second, results from a generalized $(\pi \neq 1/2)$ Bernoulli presentation schedule fall directly from Theorem 4.2 (Eq (5.8)). And finally, statistics involving response probabilities to a prticular stimulus (S_1 or S_2) are easily obtained (Eqs. (5.9) and (510)). Of course these results could be obtained without recourse to th framework, but the compatibility of the framework and the model suggets that there are dividends to be gained by an axiomatization of a moel in terms of Definition 4.5. Next, we turn to an analysis of the al-or-none multi-level model for $N = 2$.

The All-or-none Multi-level Model $(N = 2)$

Assume a list of pairs of related items for which related pairs are assigned the same response. For example,

Stimulus	Response
ABC	1
ADE	1
FGH	2
FIJ	2
KLM	3
KNO	3

might be such a list with three sublists of size 2 fitting the above criterion. In Chapter 3, the all-or-none multi-level model for learning such a list was axiomatized. The analyses of the model in Chapter 3 were restricted either to general theorems (Theorems 3.1, 3.2, 3.3) or to the special case where $c = r$ (Table 3.2). In this section a further analysis of the model in terms of the framework will be presented.

For the all-or-none multi-level model $(M = 2)$ we have $T_I = \{U,P,R\}$, $\mathcal{T} = T \times T$, and $\mathcal{N} = \{(U,R), (R,U), (P,R), (R,P)\}$, where $\mathcal{N}$ is the set of null states (Definition 4.8). Thus, the state space for the analysis is $\mathcal{J} = \mathcal{T} - \mathcal{N} = \{(U,U), (U,P), (P,U), (P,P), (R,R)\}$. $\mathcal{P} = \{\mathbb{P}_1, \mathbb{P}_2\}$, where

$$(5.11) \quad \mathbb{P}_1 = \begin{array}{c} \\ (R,R) \\ (P,P) \\ (P,U) \\ (U,P) \\ (U,U) \end{array} \begin{array}{c} \begin{array}{ccccc} (R,R) & (P,P) & (P,U) & (U,P) & (U,U) \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 & 0 \\ c & 0 & 1-c & 0 & 0 \\ r & p & 0 & 1-r-p & 0 \\ r & 0 & p & 0 & 1-r-p \end{array}\right] \end{array}$$

and

$$(5.12) \quad \mathbb{P}_2 = \begin{array}{c} \\ (R,R) \\ (P,P) \\ (P,U) \\ (U,P) \\ (U,U) \end{array} \begin{array}{c} \begin{array}{ccccc} (R,R) & (P,P) & (P,U) & (U,P) & (U,U) \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 & 0 \\ r & p & 1-r-p & 0 & 0 \\ c & 0 & 0 & 1-c & 0 \\ r & 0 & 0 & p & 1-r-p \end{array}\right] \end{array} .$$

It should be noted that the preceding state space and matrices really apply to the sublists consisting of pairs of related items, i.e., $\mathcal{S}^*$, the levels partition of $\mathcal{S}$, consists of two-item equivalence classes (Theorem 4.4). For example, the levels partition for the six-item list on p. 98 is

$$\mathcal{S}^* = \{\{ABC, ADE\}, \{FGH, FIJ\}, \{KLM, KNO\}\} .$$

The response rule $\mathcal{L}$ asserts that responses to items in state U are correct with probability g and incorrect with probability 1 - g; whereas, responses to items in states P and R are always correct. Finally, the start vector $\vec{p}_1 = (0,0,0,0,1)$.

The model, $\mathfrak{M}$, is commutative in the sense of Definition 4.7. This fact can be seen by noting

$$(5.13) \quad \mathbb{P}_1 \cdot \mathbb{P}_2 = \mathbb{P}_2 \cdot \mathbb{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1-(1-e)^2 & (1-e)^2 & 0 & 0 & 0 \\ e+(1-e)r & p(1-e) & (1-e)(1-r-p) & 0 & 0 \\ e+(1-r)r & p(1-e) & 0 & (1-e)(1-r-p) & 0 \\ r(2-r) & p^2 & p(1-r-p) & p(1-r-p) & (1-r-p)^2 \end{bmatrix} .$$

Now to apply Theorem 4.3, let us assume k_1 S_1 and k_2 S_2 presentations for the first N trials on a related pair of items $\{S_1 S_2\}$, for $k_1 + k_2 = N$. Then

$$(5.14) \quad \begin{aligned} Pr(A_{1,N+1} | S_{1,N+1}) = 1[\vec{p}^{(5)}_{N+1} + \vec{p}^{(4)}_{N+1} + \vec{p}^{(3)}_{N+1}] \\ + g[\vec{p}^{(2)}_{N+1} + \vec{p}^{(1)}_{N+1}] . \end{aligned}$$

where $\vec{p}_{N+1}^{(i)}$ is the probability of being in state i on trial $N + 1$ after the specified S_1 and S_2 presentations (we are assigning numbers to states as follows: 1-(R,R), 2-(P,P), 3-(P,U), 4-(U,P, 5-(U,U)). Now by Theorem 4.3

$$\vec{p}_{N+1} = \vec{p}_1 P_1^{k_1} \cdot P_2^{k_2}$$

$$= (0,0,0,1) \cdot \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1-(1-c)^{k_1} & (1-c)^{k_1} & 0 & 0 & 0 \\ 1-(1-c)^{k_1} & 0 & (1-c)^{k_1} & 0 & 0 \\ B_1 & A_1 & 0 & (1-r-p)^{k_1} & 0 \\ B_1 & 0 & A_1 & 0 & (1-r-p)^{k_1} \end{bmatrix} \tag{5.15}$$

$$\cdot \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1-(1-c)^{k_2} & (1-c)^{k_2} & 0 & 0 & 0 \\ B_2 & A_2 & (1-r-p)^{k_2} & 0 & 0 \\ 1-(1-c)^{k_2} & 0 & 0 & (1-c)^{k_2} & 0 \\ B_2 & 0 & 0 & A_2 & (1-r-p)^{k_2} \end{bmatrix}$$

$$= (B_1+A_1B_2) + (1-r-p)^{k_1} B_2,\ A_1A_2,\ A_1(1-r-p)^{k_2},\ A_2(1-r-p)^{k_1},$$
$$(1-r-p)^{k_1+k_2}) ,$$

where

$$A_1 = \frac{p}{r+p-c}\left((1-c)^{k_1-1} - (1-r-p)^{k_1-1}\right),$$

and

$$B_1 = 1 - A_1 - (1-r-p)^{k_1} .$$

The appropriate $\bar{p}_{N+1}^{(i)}$ terms in Eq. (5.15) can be substituted into Eq. (5.13) to obtain response probabilities as a function of k_1, k_2, c, p, r. An experiment in which the presentation orders of the S_1 and S_2 items are varied in position and in number should provide a strong test for the all-or-none multi-level model.

Next, suppose a Bernoulli presentation schedule (Definition 4.6) with $\pi = \Pr(S_1)$. Then from Theorem 4.2,

$$(5.16) \quad A_\pi = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 & 0 \\ \pi c+(1-\pi)r & (1-\pi)p & \begin{matrix} \pi(1-c) \\ + \\ (1-\pi)(1-r-p) \end{matrix} & 0 & 0 \\ \pi r+(1-\pi)c & \pi p & 0 & \begin{matrix} \pi(1-r-p) \\ + \\ (1-\pi)(1-c) \end{matrix} & 0 \\ r & 0 & \pi p & (1-\pi)p & 1-r-p \end{bmatrix} .$$

A_π could be raised to the N^{th} power to get state and response probabilities for trial N. The result will not be presented here.

The all-or-none multi-level model is a symmetric model (Definition 4.11). Hence, Theorem 4.5 can be used to lump $A_{\frac{1}{2}}$ into the states $T_1 = \{(R,R)\}$, $T_2 = \{(P,P)\}$, $T_3 = \{(P,U), (U,P)\}$, $T_4 = \{(U,U)\}$. The result is

$$(5.17) \quad A_{\frac{1}{2}} = \begin{array}{c|cccc|} & T_1 & T_2 & T_3 & T_4 \\ \hline T_1 & 1 & 0 & 0 & 0 \\ T_2 & c & 1-c & 0 & 0 \\ T_3 & \frac{1}{2}(r+c) & \frac{1}{2}p & 1-\frac{1}{2}(r+c+p) & 0 \\ T_4 & r & 0 & p & 1-r-p \end{array} .$$

The response probabilities for this chain are as follows:

$$(5.18) \qquad \Pr(x_N = 0|T_i) = \begin{cases} 1 & \text{for } T_1 \text{ and } T_2 \\ \frac{1}{2}(1+g) & \text{for } T_3 \\ g & \text{for } T_4 . \end{cases}$$

Since T_1 and T_2 are both perfect performance states in $A_{\frac{1}{2}}$, there is a simpler equivalent three-state model. This is given by

$$(5.19) \quad A'_{\frac{1}{2}} = \begin{array}{c} \\ W_1 \\ W_2 \\ W_3 \end{array} \begin{array}{c} \begin{array}{ccc} W_1 & W_2 & W_3 \end{array} \\ \left[\begin{array}{ccc} 1 & 0 & 0 \\ \frac{1}{2}(r+c+p) & 1-\frac{1}{2}(r+c+p) & 0 \\ r & p & 1-r-p \end{array}\right] \end{array} \quad \begin{array}{c} \Pr(x_n = 0|\text{row state}) \\ \left[\begin{array}{c} 1 \\ \frac{1}{2}(1+g) \\ g \end{array}\right] \end{array},$$

where $W_1 = \{T_1,T_2\}$, $W_2 = \{T_3\}$, $W_3 = \{T_4\}$. $A'_{\frac{1}{2}}$ is a two-stage model (cf. Bower and Theois, 1964). Analysis is facilitated by expanding W_2 into an error and a success state (see p. 88, Chapter 4).

$A'_{\frac{1}{2}}$ represents the three-state stochastic matrix that corresponds to the stochastic model govening the error-success process on the item pair for a Bernoulli presentation schedule with $\pi = 1/2$, i.e., each error-success protocol for the pair of items, S_1, S_2, is a sample path from this process. Thus $A'_{\frac{1}{2}}$ represents a particular stochastic process derived from the all-or-none multi-level model under the boundary conditions of $\pi = 1/2$ and the level of analysis chosen as the pair of items. Without dwelling on the point, there is a sense in which the framework provides a method for axiomatizing a theory for list learning in such a way that a particular model can be derived in accord with the boundary conditions of the experiment. This property is a feature of theories in physics, e.g., Newtonian mechanics.

An additional point can be made about a model viewed in terms of the framework. The question of whether a theory is identifiable in the sense of Greeno and Steiner (1964) can not be answered, as such, by models in the framework. The Greeno, Steiner analysis concerns the identifiability of a model for a particular presentation schedule and a particular level of analysis. Thus, a derivation, such as the model represented in Eq. (5.19) for the pair of items, provides a stochastic process (or model) which might or might not be identifiable in the sense of Greeno and Steiner. However, some additional development of the theory of identifiability is needed to apply it to a particular model, $\mathfrak{M} = (\mathcal{T}, \mathcal{P}, \mathcal{A})$. No attempt to extend identifiability in the indicated direction is presented in this paper.

Similar techniques can be used to handle the anticipation procedure. On any cycle, either the presentation order S_1S_2 or the order S_2S_1 is presented to the subject. Since the model is commutative, the effective matrix of transition probabilities for any cycle is given by Eq. (5.13). Since (R,R) and (P,P) are perfect performance states, the effective matrix on a cycle is lumpable to

$$(5.20) \qquad P_e = \begin{array}{c} \\ T_1 \\ T_2 \\ T_3 \\ T_4 \end{array} \begin{array}{c} \begin{array}{cccc} T_1 & T_2 & T_3 & T_4 \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1-(1-e)^2 & (1-e)^2 & 0 & 0 \\ e+(1-e)r & g(1-e) & (1-e)(1-r-g) & 0 \\ r(2-r) & g^2 & 2g(1-r-g) & (1-r-g)^2 \end{array}\right] \end{array},$$

where T_1, T_2, T_3, T_4 are defined in Eq. (5.17). This stochastic matrix represents the model for the analysis of the error-success subsequences associated with whichever item appears first in a cycle. Thus, if s is an error-success sequence for the item pair in an anticipation procedure, the subsequence corresponding to the even terms in s is an error-success sequence for the model in (5.20). The response probabilities, given state T_i, are presented in Eq. (5.18).

In a similar manner to the way in which Eq. (5.19) represents an equivalent model to Eq. (5.17), a three-state equivalent model (with states $W_i, i=1,2,3$) to Eq. (5.20) can be derived. The result is

$$(5.21) \qquad \mathbb{P}'_c = \begin{array}{c} \\ W_1 \\ W_2 \\ W_3 \end{array} \begin{array}{c} \begin{array}{ccc} W_1 & W_2 & W_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ (r+p)(1-c)+c & (1-c)(1-r-p) & 0 \\ r(2-r)+p^2 & 2p(1-r-p) & (1-r-p)^2 \end{bmatrix} \end{array} .$$

Computations for this model would proceed similarly to those for the model in Eq. (5.19).[5] The point of interest is that the models in Eqs. (5.19) and (5.21) are different models. Each is relevant to a different presentation procedure and each applies to a different level of analysis; however, both are derived from the all-or-none multi-level model. Next, we present a slight modification of the all-or-none multi-level model and indicate the direction of an analysis of this model in terms of the framework.

5/ It should be reiterated that models derived from Theorem 4.5 generally have differential probabilities of learning following errors and successes in a particular lumped state. This model is no exception. Analysis is facilitated by expanding (5.21) into a W_2 error state and a W_2 success state.

Another Version of the All-or-none Multi-level Model

Thus far, we have reported an example where response probability to an item depends on the states of other items in the list, and an example where items other than the one presented can change their states. For completeness we mention an extension of the all-or-none multi-level model which displays the third type of dependency discussed, namely, the transition probabilities for items may depend on the state of a particular unpresented item.

Except for one modification, the model assumes the same structure as the all-or-none multi-level model $(M = 2)$. The probability of rule learning is assumed to be e for any item presented, provided there is at least one item in the list not in state U; otherwise it is assumed to be r. For $M = 2$ the two members of $\mathbb{P}$ are displayed below:

$$
(5.22) \qquad \mathbb{P}_1 = \begin{array}{c} \\ (R,R) \\ (P,P) \\ (P,U) \\ (U,P) \\ (U,U) \end{array}
\begin{array}{c} \begin{array}{ccccc} (R,R) & (P,P) & (P,U) & (U,P) & (U,U) \end{array} \\
\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ e & 1-e & 0 & 0 & 0 \\ e & 0 & 1-e & 0 & 0 \\ e & p & 0 & 1-e-p & 0 \\ r & 0 & p & 0 & 1-r-p \end{bmatrix} \end{array} ,
$$

and

$$
(5.23) \qquad \mathbb{P}_2 = \begin{array}{c|ccccc} & (R,R) & (P,P) & (P,U) & (U,P) & (U,U) \\ \hline (R,R) & 1 & 0 & 0 & 0 & 0 \\ (P,P) & e & 1-e & 0 & 0 & 0 \\ (P,U) & e & p & 1-e-p & 0 & 0 \\ (U,P) & e & 0 & 0 & 1-e & 0 \\ (U,U) & r & 0 & 0 & p & 1-r-p \end{array} .
$$

This model has a sort of proactive feature to it in the sense that previous presentations of other items can affect the probabilities of rule learning to a particular item. The model is not commutative model. For $N = 2$, this is shown by computing $P_1 \cdot P_2$ and $P_2 \cdot P_1$. The result is

$$(5.24)\quad P_1 \cdot P_2 = \begin{array}{c} (R,R) \\ (P,P) \\ (P,U) \\ (U,P) \\ (U,U) \\ \\ \end{array} \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 1-(1-e)^2 & (1-e)^2 & 0 & 0 & 0 \\ 1-(1-e)^2 & p(1-e) & (1-e)(1-e-p) & 0 & 0 \\ 1-(1-e)^2 & p(1-e) & 0 & (1-e)(1-e-p) & 0 \\ 1-(1-r)^2 & p^2 & p(1-e-r) & p(1-r-p) & (1-r-p)^2 \\ + \; p(e-r) & & & & \end{array}\right],$$

and

$$(5.25)\quad P_2 \cdot P_1 = \begin{array}{c} (R,R) \\ (P,P) \\ (P,U) \\ (U,P) \\ (U,U) \\ \\ \end{array} \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 1-(1-e)^2 & (1-e)^2 & 0 & 0 & 0 \\ 1-(1-e)^2 & p(1-e) & (1-e)(1-e-p) & 0 & 0 \\ 1-(1-e)^2 & p(1-e) & 0 & (1-e)(1-e-p) & 0 \\ 1-(1-r)^2 & p^2 & p(1-r-p) & p(1-e-p) & (1-r-p)^2 \\ + \; p(e-r) & & & & \end{array}\right].$$

These two matrices differ in their (5,3) and (5,4) terms.

This model, in miniature form, embodies some of the ideas currently being worked on by G. Groen and L. Hyman (personal communication). They are investigating the assumption that the probability a concept is learned on any trial depends on the number of items in the list that have been learned as paired associates. The above model reflects these considerations by setting the concept learning parameter equal to one value if no items have been learned as paired associates and a second value if any items have been so learned. Further analysis of this model will not be presented in this paper. With the exception of the lack of commutativity, the analysis would proceed along the same lines as the all-or-none multi-level model. Next we turn to an analysis of Restle's strategy-selection theory within the framework of Chapter 4.

Strategy Selection Theory

Restle's strategy-selection theory (Restle, 1962, 1964; Polson, Restle, Polson, 1965) has been mentioned in Chapter 4, p. 44 and p. 65. In this section we present one possible interpretation of his theory in terms of the framework. As will be seen, there are two reasons why his theory is an attractive one to analyze by our methods. The first is that it provides a complement to the all-or-none multi-level model. The multi-level model has the property that similar stimuli are paired with the same response; whereas, in strategy-selection applications, similar stimuli are paired with different responses. Thus, stimulus confusion facilitates performance in the former situation and hinders it in the latter. The second attraction to analyzing Restle's theory in terms of the framework comes from noting that in Restle's applications of his theory several approximations are made (Restle, 1964, pp. 132-144; Polson, Restle, Polson, 1965). Restle is aware of these approximations and even suggests that a completely accurate analysis of his theory would require a complicated Markov chain analysis involving the whole set of confusable items and different transition matrices for different items (Restle, 1964, pp. 168-171). The method of dealing with dependent items (in this case confusable ones) in the framework appears to be similar to the method of analysis Restle had in mind.

Restle applies strategy-selection theory to a number of experiments. In applications, the theory takes the form of a finite state Markov chain. The intermediate states of the model involve stimulus confusion or response confusion. Since many of his applications are similar, the main points of this section can be made in the context of the Polson, Restle, Polson (1965) experiment. Next, we turn to a description of the experiment and model reported in that paper.

In the experiment, college students learned a 16-item paired-associate list with 5 response alternatives by the anticipation procedure. The stimuli were symbols such as a chess knight, a question mark, and musical notes. The responses were common four-letter words. The major manipulation was that 8 of the items were highly dissimilar; whereas, the other 8 items consisted of 4 highly-confusable pairs, e.g., two very similar Chinese words. Confusable stimuli were assigned different responses.

The model assumed by Polson, Restle, and Polson had the property that unique (non-confusable) S-R pairs would be learned by a two-stage all-or-none model (the one-element P-level model); whereas, the confusable twin items would be learned by a three-stage model. The intermediate stage was a stimulus confusion stage. More specifically, the model for confusable pairs assumes three states, S_U, S_I, S_L, where S_U is an initial unlearned state with correct responses emitted with probability p, S_I is an intermediate confusion state where correct responses are made with probability F and confusion responses (incorrect responses which would be correct for the twin) with probability G, and S_L is a final learned state. The transition matrix for the model is as follows:

$$(5.26)\qquad \begin{array}{c} \\ S_{L,n} \\ S_{I,n} \\ S_{U,n} \end{array} \begin{array}{c} \begin{array}{ccc} S_{L,n+1} & S_{I,n+1} & S_{U,n+1} \end{array} \\ \left[\begin{array}{ccc} 1 & 0 & 0 \\ Gd & 1-Gd & 0 \\ cd & c(1-d) & 1-c \end{array}\right] \end{array} \qquad \begin{array}{c} \Pr(\text{correct}\,|\,\text{row state}) \\ \left[\begin{array}{c} 1 \\ F \\ p \end{array}\right] \end{array},$$

where it is understood that transitions take place from S_I to S_L only on confusion errors. Thus, c is the probability any strategy is selected to an item in state S_U, and d is the probability a selected strategy is not a confusion one. Resampling of strategies is assumed to take place only on errors. The model in Eq. (5.26) is assumed to govern the learning of a single confusable item, i.e., the model was applied to a P-level analysis of twinned items in the Polson, Restle, Polson paper.

There are several reason why Eq. (5.26) does not adequately embody some features of strategy-selection theory. To see these reasons, it will be helpful to rewrite the model by expanding the intermediate S_I state into an intermediate error state, S_I^-, and an intermediate success state, S_I^+. The result is

$$
(5.27)\qquad
\begin{array}{c}
 \\ S_L \\ S_I^e \\ S_I^c \\ S_0
\end{array}
\begin{array}{c}
\begin{array}{cccc} S_L & S_I^e & S_I^c & S_0 \end{array} \\
\left[\begin{array}{cccc}
1 & 0 & 0 & 0 \\
d & (1-d)Q & (1-d)P & 0 \\
0 & Q & P & 0 \\
cd & c(1-d)Q & c(1-d)P & 1-c
\end{array}\right]
\end{array}
\qquad
\begin{array}{c}
\Pr(\text{correct}\,|\,\text{row state}) \\
\left[\begin{array}{c} 1 \\ 0 \\ 1 \\ P \end{array}\right].
\end{array}
$$

Strategy-selection theory postulates that once a strategy has been sampled, resampling occurs only on an error. Careful analysis reveals that Eq. (5.27) does not represent this assumption in a way that keeps harmony with the theory. To see this point, let S_1 and S_2 denote a pair of confusable stimuli. Suppose a confusion strategy, h_1, is learned when S_1 appears. By its nature h_1 will produce correct responses to S_1 and errors to S_2.[6] Since h_1 was learned when S_1 appeared, the subject is now in state S_I for item S_1; however, only on a trial when S_2 appears, h_1 is tried with an error, and resampling occurs, will a transition take place from S_I. The error that causes rejection of h_1 does not take place on a trial when S_1 is presented but on a trial when S_2 appears. But (5.27) assumes that each subject-item protocol is a sample path from this learning-only-on-errors model. The error that causes learning is not in the protocol for S_1; and, thus, learning can take place following a success to S_1 if an intermediate S_2 item causes rejection of the confusion strategy learned when S_1 was previously presented.

---

[6] The reader who doubts that our treatment of strategy-selection is a fair interpretation is referred to Restle (1964), pp. 126-127. Actually, it is this stimulus-specific interpretation of strategy sampling that this writer finds so attractive about Restle's theory.

A possible way to rectify the state of affairs might be to use the model represented in Eq. (5.27) to account for the error-success process on the pair $\{S_1, S_2\}$, i.e., the level of analysis for which pairs of items are the units. Restle (1964, p. 125) suggests this by arguing that when stimulus generalization is considered, "the unit of analysis must be the subset of related items as learned by a single subject." If the unit of analysis for Eq. (5.27) is the item pair, the learning-only-on-errors assumption is no longer in question. However, suppose s_1 is learned on a trial when S_1 appears and is rejected when S_2 appears in favor of a strategy which is unique to S_2. What strategy now covers S_1? The answer is that S_1 is thrust back to the unlearned state, S_U. But this has zero probability in Eq. (5.27). Polson, Restle, Polson (1965, p. 54) point out this possibility and even note properties of the data to indicate that such events did happen in their experiment.

One resolution to these problems would be to change the transition probabilities in Eq. (5.27). This solution seems not to be desirable since the model already fails to reflect the nature of the intra-pair dependencies postulated by strategy-selection theory. A better resolution would be to attempt to embody these dependencies in a multi-level model written in terms of the framework. This direction is very definitely suggested by Restle (1964, pp. 168-171). One possible model embodying strategy-selection assumptions for the Polson, Restle, Polson experiment is presented next.

Suppose the item state space, T_I, is $\{U, C_1, C_2, L\}$, where U is an unlearned state, C_i is a state where a confusion strategy requiring response i is held (for i = 1,2), and L is a learned state. After removing null states, the state space for the list is as follows:

$$\mathcal{S} = \{(U,U),\ (C_1,C_1),\ (C_2,C_2),\ (U,L),\ (L,U),\ (L,L)\}\ .$$

$\mathcal{P}$ consists of two matrices, P_1 and P_2; they are as follows

$$(5.28)\ P_1 = \begin{array}{c} \\ (L,L) \\ (L,U) \\ (U,L) \\ (C_2,C_2) \\ (C_1,C_1) \\ (U,U) \end{array} \begin{array}{c} \begin{array}{cccccc} (L,L) & (L,U) & (U,L) & (C_2,C_2) & (C_1,C_1) & (U,U) \end{array} \\ \left[\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ c & 0 & 1-c & 0 & 0 & 0 \\ 0 & d & 0 & 0 & 1-d & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & cd & 0 & 0 & c(1-d) & (1-c) \end{array}\right] \end{array},$$

and

$$(5.29)\ P_2 = \begin{array}{c} \\ (L,L) \\ (L,U) \\ (U,L) \\ (C_2,C_2) \\ (C_1,C_1) \\ (U,U) \end{array} \begin{array}{c} \begin{array}{cccccc} (L,L) & (L,U) & (U,L) & (C_2,C_2) & (C_1,C_1) & (U,U) \end{array} \\ \left[\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \\ c & 1-c & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & d & 1-d & 0 & 0 \\ 0 & 0 & cd & c(1-d) & 0 & 1-c \end{array}\right] \end{array},$$

where the following special assumptions have been made: (1) If an item is presented and a confusion strategy is learned, it applies equally to both items if they were previously unlearned, (2) if one item is learned and the other is not, any strategy learned on a trial when the unlearned item is presented is sufficient to move the pair into state (L,L), and (3) on an error trial to a confusion strategy, only the presented item can be learned and, if so, its twin goes to state U. These assumptions appear to be in the spirit of strategy-selection theory but, by no means, represent the only way strategy-selection theory could be formalized in terms of the framework. The response rule, $\mathcal{L}$, would specify that items in U would be responded to correctly with probability p, in L with probability 1, and in state C_1C_1 the response correct for S_1 is always made.

$\mathfrak{M} = (\mathcal{J}, \mathcal{P}, \mathcal{L})$ is not a commutative model, since $\mathbb{P}_1 \cdot \mathbb{P}_2 \neq \mathbb{P}_2 \cdot \mathbb{P}_1$. The results of the matrix multiplications are as follows:

$$(5.30)\quad \mathbb{P}_1 \cdot \mathbb{P}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ e & 1-e & 0 & 0 & 0 & 0 \\ e & 0 & 1-e & 0 & 0 & 0 \\ ed & d(1-e) & d(1-d) & (1-d)^2 & 0 & 0 \\ 0 & 0 & d & 1-d & 0 & 0 \\ e^2d & ed(1-e) & ed(2-d-e) & (1-d)e(2-e-d) & 0 & (1-e)^2 \end{bmatrix},$$

$$(5.31)\quad \mathbb{P}_2 \cdot \mathbb{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ e & 1-e & 0 & 0 & 0 & 0 \\ e & 0 & 1-e & 0 & 0 & 0 \\ 0 & d & 0 & 0 & 1-d & 0 \\ ed & (1-d)d & d(1-e) & 0 & (1-d)^2 & 0 \\ e^2d & ed(2-e-d) & ed(1-e) & 0 & e(1-d)(2-e-d) & (1-e)^2 \end{bmatrix}.$$

The anticipation procedure requires that the two possible orders of presentation, S_1S_2 and S_2S_1, are equally likely. In order to apply

Theorem 4.2 we compute the average effective matrix, $\mathbb{A}$, on a cycle (see Theorem 4.2). The result is

$$(5.32)\quad \mathbb{A} = \tfrac{1}{2}\,\mathbb{P}_1 \cdot \mathbb{P}_2 + \tfrac{1}{2}\,\mathbb{P}_2 \cdot \mathbb{P}_1$$

$$= \begin{array}{c} \\ (L,L) \\ (L,U) \\ (U,L) \\ (C_2,C_2) \\ (C_1,C_1) \\ (U,U) \end{array} \begin{array}{cccccc} (L,L) & (L,U) & (U,L) & (C_2,C_2) & (C_1,C_1) & (U,U) \end{array} \atop \left[\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \\ e & 1-e & 0 & 0 & 0 & 0 \\ e & 0 & 1-e & 0 & 0 & 0 \\ \frac{ed}{2} & \frac{d(2-e)}{2} & \frac{d(1-d)}{2} & \frac{(1-d)^2}{2} & \frac{(1-d)}{2} & 0 \\ \frac{ed}{2} & \frac{d(1-d)}{2} & \frac{d(2-e)}{2} & \frac{(1-d)}{2} & \frac{(1-d)^2}{2} & 0 \\ e^2d & \frac{ed}{2}(3-2e-d) & \frac{ed}{2}(3-2e-d) & \frac{e(1-d)(2-e-d)}{2} & \frac{e(1-d)(2-e-d)}{2} & (1-e)^2 \end{array}\right].$$

Since $\mathfrak{M}$ is a symmetric model, Theorem 4.5 can be used to lump $\mathbb{A}$ to a four-state matrix with states $T_1 = \{(L,L)\}$, $T_2 = \{(L,U), (U,L)\}$, $T_3 = \{(C_2,C_2), (C_1,C_1)\}$, $T_4 = \{(U,U)\}$. The result is

$$(5.33)\qquad \mathbb{A}' = \begin{array}{c|cccc} & T_1 & T_2 & T_3 & T_4 \\ \hline T_1 & 1 & 0 & 0 & 0 \\ T_2 & e & 1-e & 0 & 0 \\ T_3 & \frac{ed}{2} & \frac{d}{2}(3-e-d) & \frac{(1-d)(2-d)}{2} & 0 \\ T_4 & e^2d & ed(3-2e-d) & e(1-d)(2-e-d) & (1-e)^2 \end{array}.$$

The probability of correct given state T_i is as follows:

$$(5.34) \qquad Pr(x_n = 0 \mid \overline{T}_i) = \begin{cases} 1 & \text{if } T_1 \\ \frac{1}{2}(1-g) & \text{if } T_2 \\ \frac{1}{2} & \text{if } T_3 \\ g & \text{if } T_4 \end{cases}$$

The model represented by Eqs. (5.33), (5.34) would apply to the error-success sequences on items S_1 or S_2, which appear first on each cycle (see p.105 of this chapter for a further description of this level of analysis). This is because between two successive first appearances, each of the matrices, $\mathbb{P}_1 \cdot \mathbb{P}_2$ or $\mathbb{P}_2 \cdot \mathbb{P}_1$, is equally likely to be effective. Restle assumes items start in state U, so $\overline{p}_1 = (0,0,0,1)$ for this model. Since the data for first-appearing items is not presented in Polson, Restle, Polson, no attempt will be made to present statistics for this model. It should yield to hand computations of some statistics, or it could be analyzed by computer, using Bernbach's (1966) scheme. Intuition suggests that the pattern of predictions for this model should conform as well or better to data as the model presented in Polson, Restle, Polson. There are two reasons for this intuition: (1) items can drop from a confusion state to state U, and there are indications in the data that this happened, and (2) the model is an average of a convolution of two geometric distributions and a convolution of three geometric distributions. Since a convolution of two geometrics does not do badly, it is unlikely that the addition of another stage will hurt prediction. The case is not, however, entirely transparent.

The model for the error-success process on the second-appearing item in a cycle is slightly more complicated. This is because this item is always different from the item associated with the last effective matrix, i.e., if S_2 appears second on some cycle, then $\mathbb{P}_1$, corresponding to S_1, which appeared first, is the last effective matrix. If one uses the average matrix, $\mathbb{A}$, as in Eq. (5.33), it is assumed not only that $\mathbb{P}_1 \cdot \mathbb{P}_2$ and $\mathbb{P}_2 \cdot \mathbb{P}_1$ are equally likely, but also that S_1 and S_2 appear equally likely and independent of whether $\mathbb{P}_1 \cdot \mathbb{P}_2$ or $\mathbb{P}_2 \cdot \mathbb{P}_1$ was effective. This assumption is violated for second-appearing items but not for first-appearing items on a cycle. There are several ways second-appearing items can be handled, but the details will not be presented here. One way would be to consider the arrangements $\mathbb{P}_1 \mathbb{P}_2 S_1$ and $\mathbb{P}_2 \mathbb{P}_1 S_2$, which are the two possibilities for effective matrix and item-presentation for second-appearing items. By incorporating the presented item into the state space (e.g., a state might be (U, L, S_1)), a model for second-appearing items could be derived.

Additional results and statistics for different presentation schedules and levels of analysis could be presented for strategy-selection theory as interpretated by the framework. These will not be presented in this paper. It is hoped that this section has indicated the direction that a mathematical theory for confusion processes in list learning might take. This section concludes our analysis of models in terms of the framework. We have seen how the theorems of Chapter 4 can be applied to a variety of multi-level models embodying various sorts of item dependencies. The net value of the framework depends entirely on its ability to generate new and tractable tests for learning models.

CHAPTER 6

EXPERIMENTS AND CONCLUSIONS

In the first part of this chapter we will discuss two experiments that the writer has conducted to generate some data relevant to the ideas and methods of analyses discussed in Chapters 2 and 3. Since these experiments represent only the start of a program to pursue experimentally the ideas in those chapters, their presentation has been postponed to this last chapter, which is designed to indicate plans for developing and extending the ideas in this paper. In the last part of the chapter we will indicate briefly some general directions that research motivated by the ideas in Chapters 4 and 5 might take.

Experiments

Before presenting the two experiments, it will be useful to describe the general paradigm that governs the design of both. The paradigm involves list learning. The stimulus terms are composed of recognizable components with some number N of these components per stimulus (in the experiments to be reported, $N = 3$). There are fewer response terms than stimulus terms, and hence, more than one stimulus is paired with each response.

Some of the components making up a stimulus are unique in the sense that they only appear as components of that stimulus, whereas other components are shared by more than one stimulus. The major manipulation in the paradigm is to construct stimuli and assign responses in such a way that all stimuli sharing any component (or components) are paired with the same response. Thus, shared ("overlap") components should aid the subject in the sense that they will never lead him astray in his responses, i.e., if the subject pairs a certain component x to response A and hence says response A to any stimulus having component x, he will always be correct. The following is a possible structure of a typical list used in the experiments to be reported:

Stimulus			Response
Carl	Stan	Eric	1
Carl	Dave	Robert	1
Carl	George	Jim	1
Jack	Bill	Bob	1
Jerry	Dick	Pat	3
Jerry	Frank	Louis	3
Jerry	Mike	Guy	3
Tom	Harry	Glen	3

etc.

It should be noted that the only overlap components are Carl and Jerry; and, further, if the subject pairs any component with a number response, he will get the stimulus having that component correct as well as any other stimulus (if any) sharing that component.

The list structure for this paradigm is similar to that frequently employed to study concept identification (e.g., Atkinson, Bower, and Crothers, 1965, p. 31); however, there is one essential difference in the two paradigms. This difference is that overlap components in a concept identification task are not always facilitative; that is, two stimuli can share a component and yet be assigned different responses. Our paradigm is even more different from that employed by Polson, Restle, and Polson (1965) to study confusion processes in paired-associate learning. In their study, stimuli sharing common components were always assigned different responses (see pp. 109-110 for a discussion of their paradigm).

By using the paradigm described in this chapter, it was hoped that positive inter-item transfer within the list would result from the facilitative nature of the overlap components. As will be seen, this expectancy was borne out in the data. Additional motivations for the experiments were to gather data relevant to the levels analyses discussed in Chapter 2 and possibly to fit the all-or-none multi-level model to these data. However, only some of these latter expectancies materialized. Next, we turn to a discussion of the two experiments.

Experiment I

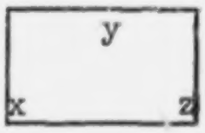
Method

Subjects.--The Ss were 15 male and female undergraduate and non-psychology graduate students at Stanford University. Each S was paid $1.50 for his participation in the experiment. The data for all Ss were used. The initial plan was to run 50 Ss in the experiment; however, the task proved so easy that only certain statistics, requiring many less than 50 Ss for stability, were usable.

Apparatus and Materials.--Subjects were run one at a time. Presentation was by hand. The E sat facing the S behind a 1 x 2 ft. screen and placed 3 x 5 inch cards on a 3 x 8 inch metal card rack situated to the E's right of the screen.

The materials consisted of three decks of twelve stimulus cards.

Each card in the experiment was composed of three component words arranged in a triangular fashion on a card, i.e., if x,y,z were the three components, a typical arrangement on a card might be



.

The responses for a particular deck were either the numbers $\{3,5,7,9\}$ or the numbers $\{2,4,6,8\}$.

The twelve stimulus cards in each deck were partitioned into four sets of three stimuli per set. Each set was assigned to a different one of the four response numbers. Each set of three stimuli in the experiment had one of the following three structures: (1) all three stimuli shared exactly one common component word, (2) two of the three stimuli shared a common component word, and (3) none of the stimuli shared a component word. Denote these three structures by C_3, C_2, and C_0, respectively. With the exception of the overlap components possible in a C_3 or C_2 structure, all other components for a particular deck were unique, i.e., appeared only on a single stimulus.

Deck (list) one consisted of animal names as the components, e.g. toad, mole, badger, and consisted of 2 C_3 and 2 C_0 sets. Lists two and three had the following structure. One of the lists had a 2 C_3, 1 C_2, 1 C_0 structure, and the other list had a 1 C_3, 2 C_2, 1 C_0 structure. The components for a particular one of these lists were either all common, short, boys' first names, e.g., Jim, Bill, Dick, or common, short, girls' first names, e.g., Patty, Ann, Margie. Each of the two orders for presenting the two lists was given to half the Ss.

Procedure.--Each S received training trials on each of the three lists. Presentation was by the paired-associate anticipation method. The inter-item interval was short, with a mean of about 1 sec. (range about .5 to 1.5 sec.). The break between cycles (randomizations) was noticeable and about 5 sec., and the break between lists was about 2 minutes. For the first list, Ss were run either to a criterion of one errorless cycle through the list or 8 complete cycles -- whichever occurred first; however, for lists II and III, they were run to a criterion of two errorless cycles. Upon the presentation of a particular stimulus card, the S, at his leisure, gave orally one of the four number responses; immediately thereafter the E told him the correct number for that stimulus.

The arrangement of components on a card was counterbalanced, both for a single S and from S to S. Within a given cycle through the list, an overlap component never appeared twice in the same position (this was accomplished by having three randomizations of each list available to the E). Finally, to further minimize recognition of the overlap components, presentation orders were arranged in such a way that two stimuli sharing a common component never appeared adjacent in a cycle. Ss were given brief paired-associate instructions and were told that the spatial arrangement of a particular set of component words on a card might change from cycle to cycle. Following the third list the S was given a paper and pencil task to see how many of the component-number pairings he could remember. The S was required to fill a response number in the blank opposite each component word.

Results and Discussion

This section will present results for List I (2 C_3 and 2 C_0 sets) first, followed by the analysis of Lists II and III. The major results for List I can be seen from a P-level analysis of the data using some of the statistics discussed in Chapter 2 (see pp. 16-18). By way of preview, these statistics are as follows: (1) the learning curve, $Pr(x_n = 1)$; (2) the mean total errors, $E(T)$, and the mean trial number of the last error, $E(L)$; and the probability distributions of these two statistics; and (3) the probability of an error prior to the last error, $Pr(x_n = 1|L > n)$, and the probability of an error given error curve, $Pr(x_{n+1} = 1|x_n = 1)$. These three classes of statistics are presented for the C_3 stimulus sets and C_0 stimulus sets separately. Figure 6.1 presents $Pr(x_n = 1)$, Table 6.1 presents $E(T)$ and $E(L)$, Fig. 6.2 presents the distributions of T and L; and Figs. 6.3, 6.4 present $Pr(x_{n+1} = 1|x_n = 1)$ and $Pr(x_n = 1|L > n)$. It should be reiterated that these statistics are computed for a P-level analysis.

First, it is quite evident from the learning curve (Fig. 6.1) and from the mean total errors and mean trial number of the last error (Table 6.1) that C_3 stimuli (stimuli with an overlap component) were learned more rapidly than C_0 stimuli. Also, there is evidence that the process governing C_3 learning produced qualitatively different data from the data for C_0. In Fig. 6.1, the C_3 learning curve is not badly fit by an exponential function; however, the learning curve for C_0 stimuli is more S-shaped. This difference could reflect the fact that Ss learned to recognize and attend to the overlap components to the detriment of stimuli in C_0 sets not having these components. A qualitative difference

Table 6.1. Mean Total Errors, E(T), and Mean Trial Number of Last Error, E(L), for C_3 and C_0 (List I).

	C_3	C_0
E(T)	2.72	3.72
E(L)	3.57	5.25

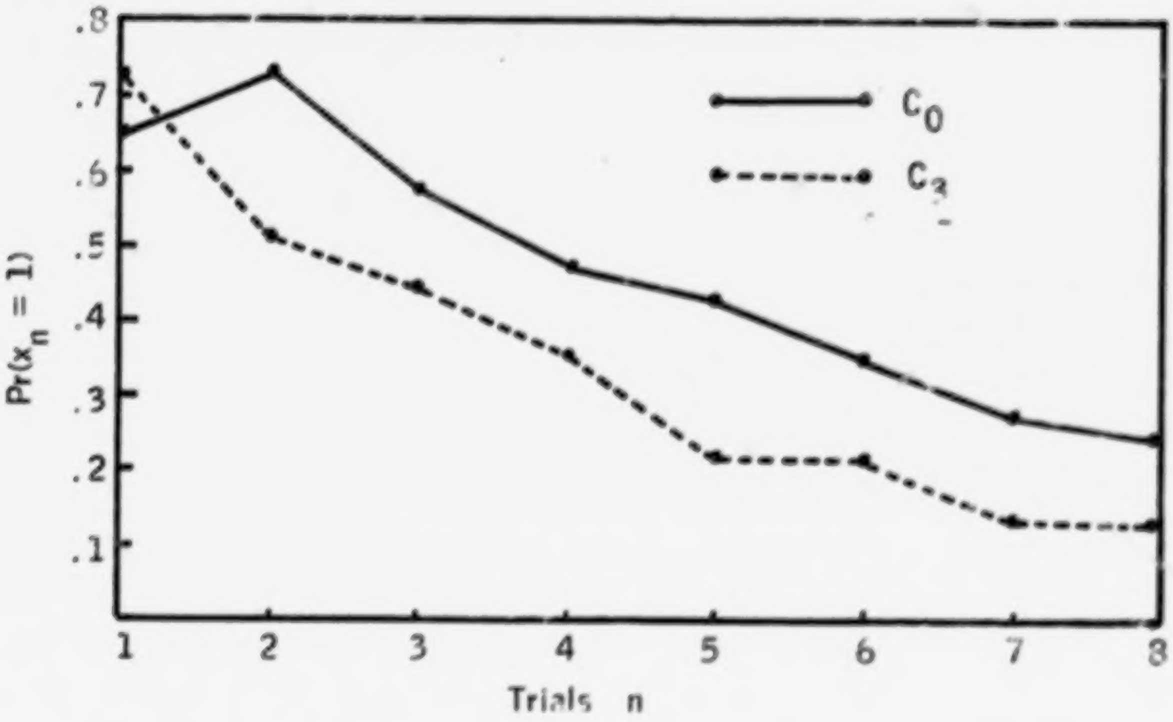


Fig. 6.1. P-level Learning Curves for C_0 and C_3, List I.

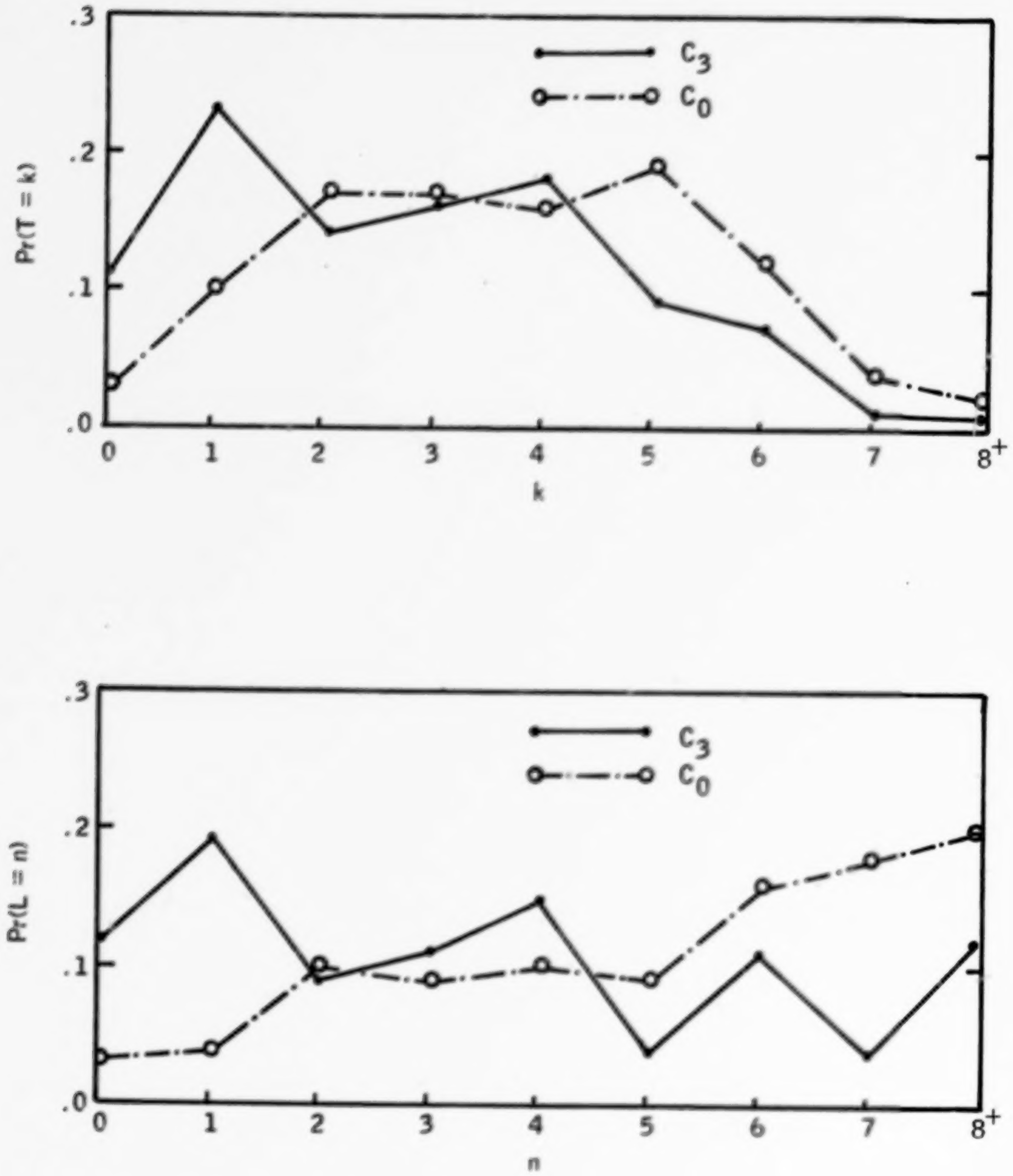


Fig. 6.2. Total Error Distributions, T, and Trial Number of Last Error Distributions, L, for C_3, C_0, List I.

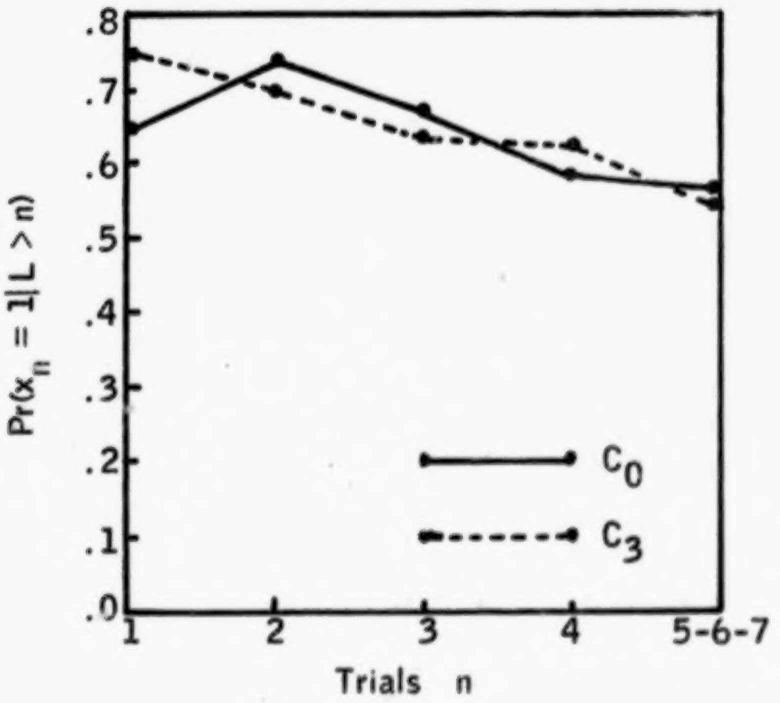


Fig. 6.3

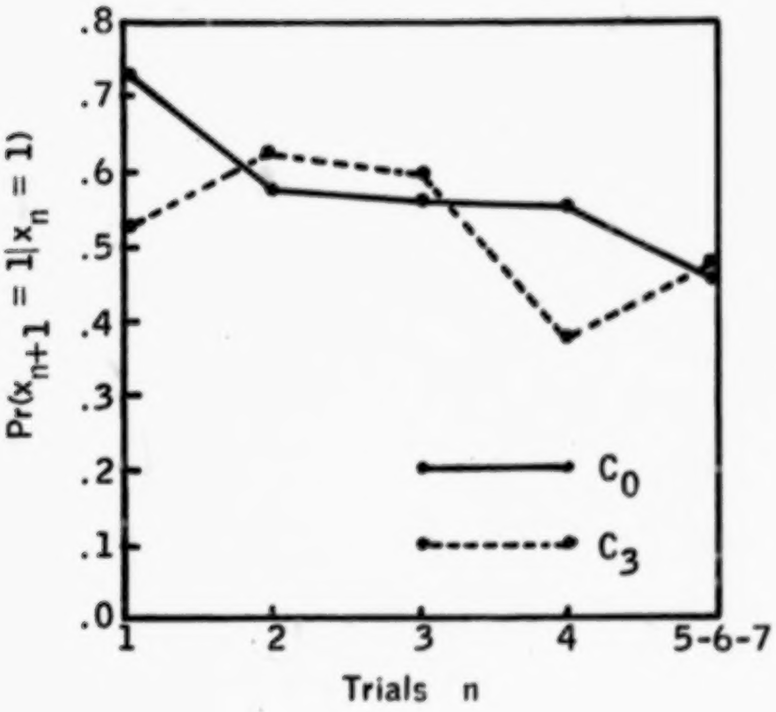


Fig. 6.4

Figs. 6.3 and 6.4. Pr(error on n prior to last error) and Pr(error on n + 1|error on n), respectively for C_3 and C_0 (List I). Data are based on from 83 to 33 cases (mostly more than 50).

in the data for C_3 and C_0 is also seen in the total error and trial number of the last error distributions (Fig. 6.2). The C_3 distribution appears somewhat geometric, although limited data prevent a sharp resolution of this point. On the other hand, the C_0 distributions are definitely not geometric.

Thus far, it appears that for C_0 we can reject the one-element P or R level model and also the all-or-none multi-level model, since these models all predict exponential learning curves and geometric T and L distributions for the P-level of data analysis (see Tables 2.1 and 3.2, p. 18 and p. 39 , respectively). Moreover, the $\Pr(x_n = 1|L > n)$ and $\Pr(x_{n+1} = 1|x_n = 1)$ curves (Figs. 6.3 and 6.4) make it unlikely that any of these three models could account for C_3 data. Both curves tend to decrease over trials, whereas all three models predict that they should be flat. Thus, it appears that processes more complicated than all-or-none P and R level mechanisms are needed to account for the data from List I.

The picture becomes more complicated in light of the R-level analyses. None of these analyses (which will not be given in detail here) revealed anything approaching a significant tendency for R-level learning (in the sense of Chapter 2) for C_3 stimuli. The R-level learning curve was essentially flat within a cycle and the P-level error-success protocols for C_3 showed no notable intercorrelations (see p. 20 and p. 21, Chapter 2). This lack of R-level learning could be reflected in the rapid learning of C_3 stimuli. Thus the S̲ might not have had a chance before reaching criterion to manifest significant transfer effects by these analyses. However, the difference in learning rate of

C_3 and C_0 stimuli strongly indicates that the overlap components were effective in cutting down errors to C_3 stimuli.

List I was designed as a warm-up task for Lists II and III. It was hoped that the S would have a fair idea of the structure of the stimulus classes after his encounter with List I, and would therefore perform in a stable fashion on Lists II and III. Next, we move to an analysis of these two lists.

Apparently there was no significant difference in the learning rate between Lists II and III (the numbers refer to the list the S saw 2^{nd}, 3^{rd}; the two lists are discussed on p. 121). Nor was there any tendency to learn the list having structure 2 C_3, 1 C_2, 1 C_0 any faster or slower than the 1 C_3, 2 C_2, 1 C_0 list. There was, however, a slight tendency to learn stimuli with boys'names as components slightly slower than stimuli with girls' names. Since the component type was randomized both for list order and list type, the data from Lists II and III were combined for analysis despite this slight differential learning rate on component type. All C_3 stimuli, all C_2 stimuli sharing a component (i.e., C_2^+ stimuli), all C_2 stimuli with all unique components (i.e., C_2^- stimuli), and all C_0 stimuli were pooled into four classes for a P-level analysis. These classes had 135, 90, 45, and 90 protocols in each class, respectively.

The P-level learning curves for the four classes are presented in Fig. 6.5 and the mean total errors and mean trial number of the last error are presented in Table 6.2. Finally, the distribution of the total error statistic is presented in Fig. 6.6. Learning was so rapid for Lists II and III that $Pr(x_n = 1|L > n)$ and $Pr(x_{n+1} = 1|x_n = 1)$ were not

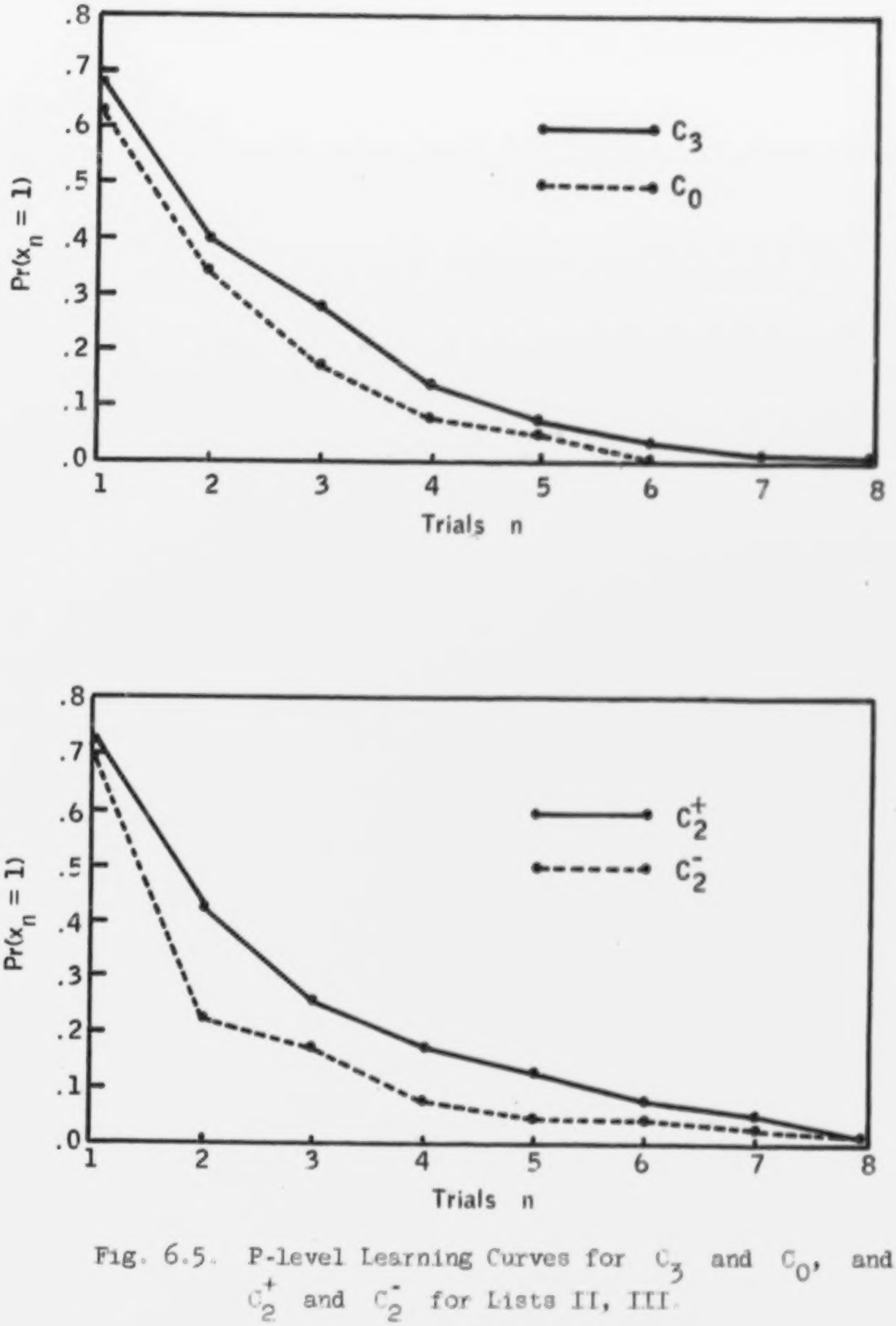


Fig. 6.5. P-level Learning Curves for C_3 and C_0, and C_2^+ and C_2^- for Lists II, III.

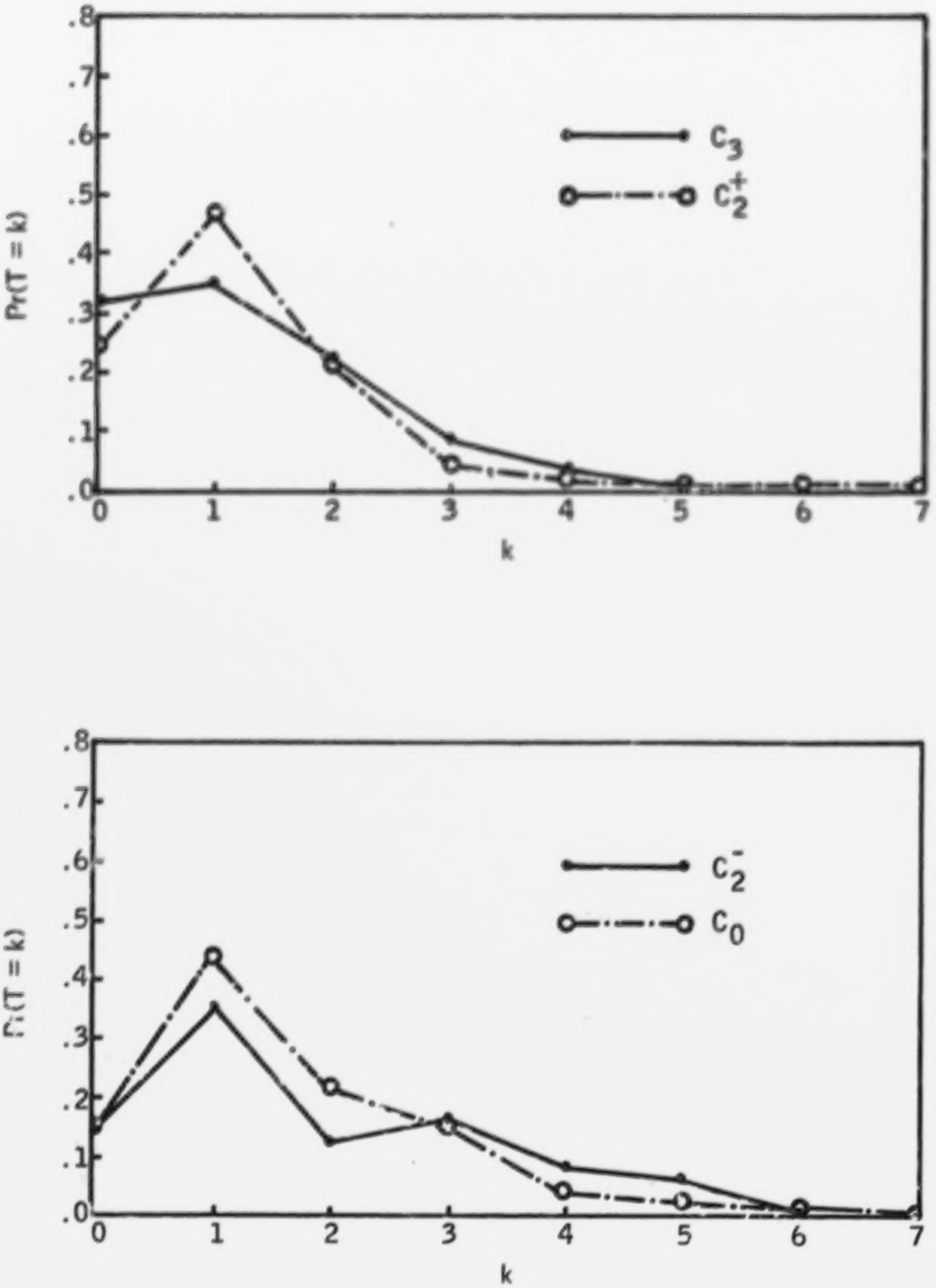


Fig. 6.6. Total Error Distributions for C_3, C_2^+, C_2^-, C_0 for Lists II, III.

Table 6.2. Mean Total Errors, $E(T)$, and Mean Trial Number of Last Error, $E(L)$, for C_3, C_2^+, C_2^-, and C_0 (Lists II and III).

	C_3	C_2^+	C_2^-	C_0
$E(T)$	1.21	1.30	1.84	1.62
$E(L)$	1.46	1.62	2.22	2.07

sufficiently stable to warrant their inclusion. Other things being equal, these two statistics tended to decrease over trials.

The learning curve analysis (Fig. 6.5) reveals at least two things. First, stimuli with overlap components were learned significantly faster than stimuli without these components. Second, learning was very rapid in the experiment with only about 15% or less errors per trial on and beyond trial 3. Closer analysis reveals that the C_3 and C_2^+ curves drop faster than an exponential function during early trials. This can be seen since the first decrement in the error probability was greater than 50%, whereas later decrements tended to be less than 50%. The C_0 and C_2^- curves are more closely approximated by an exponential function. It seems possible that the overlap components were both identified and paired with responses on the first cycle, whereas they were already identified for later cycles and possibly ignored by some Ss. Interviews did indicate some conscious ignoring of overlap components by some Ss. A section of the R-level learning curve to follow (Fig. 6.7) bears on this recognition and pairing hypothesis.

The fact that learning was quite rapid for these two 12-item lists is even more strikingly seen in Table 6.2. The mean total errors for each class was less than two. The total error distributions in Fig. 6.6 reveal that, in each of the four cases, geometric-like distributions are obtained; however, rapid learning and small N make it difficult to discriminate between a geometric distribution and one that just drops as k increases. These distributions reveal the differential difficulty in C_3 and C_2^+ vs. C_0 and C_2^- classes. The fact that $Pr(T = 0)$ is greater for C_3 than for C_2^+ might indicate more transfer from stimulus to stimulus during cycle 1 when three stimuli share a common component as opposed to just two. This transfer within cycle 1 is illustrated in the R-level learning curve to be presented later (Fig. 6.7).

A comparison of the overlap classes and non-overlap classes (C_3, C_2 vs. C_0, C_2^-) both on their total error distributions (Fig. 6.6) and their learning curves (Fig. 6.5) indicate the nature of the learning-to-learn effects developed in the experiment. The List I data indicate that trial 1 had little direct effect on C_0 stimuli, whereas trial 1 had the biggest effect on cutting down errors to C_0 stimuli for Lists II and III. Also, the C_0 total error distribution is definitely not geometric for List I and apparently geometric-like for Lists II and III. These differences are attributed to the Ss' increased familiarity with the paradigm for Lists II and III, i.e., the S learned to expect some but not all overlap components and to use them. The post-list III recall task indicated that Ss remembered the component response pairing for 85% of the overlap (relevant) components and only about 35% of the irrelevant components (corrected for guessing). Since it was necessary to learn a minimum of 18% of the irrelevant components to master the list, this measure indicates that not too much learning above the minimum necessary took place.

Another difference between Lists I vs. II and III is revealed by the R-level analysis. The small number of Ss and few errors prohibit a full R-level analysis; however, there were significantly fewer errors made to the 3rd appearing C_3 stimulus on cycle 1 than were made to the 1st and 2nd stimulus in a C_3 class. This fact is shown in Fig. 6.7, which presents a section of the R-level learning curve corresponding to

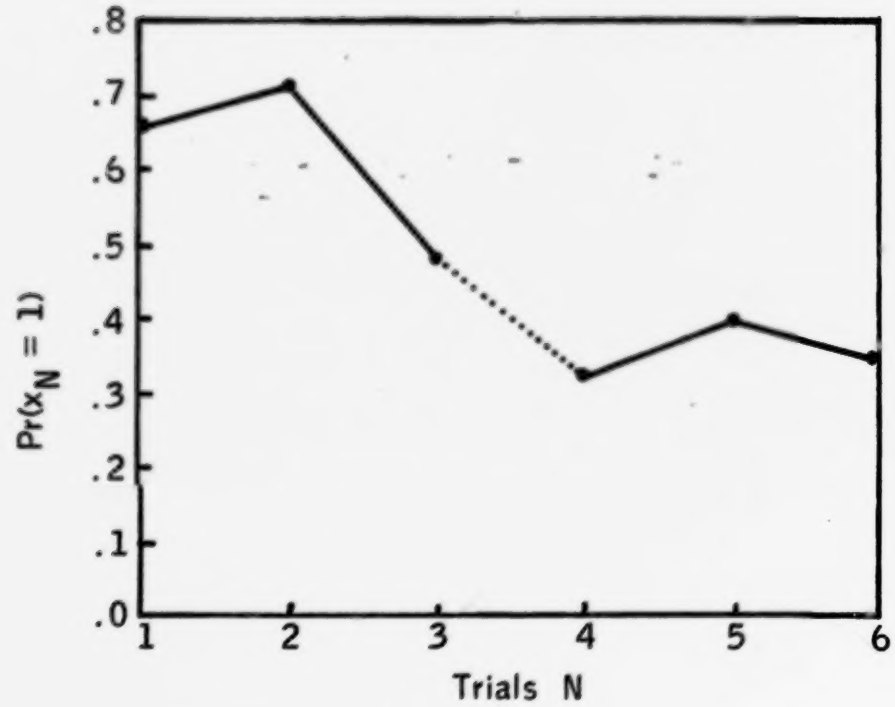


Fig. 6.7. Section of the R-level Learning Curve for C_3 for Lists II, III. The solid line is within a cycle, and the dotted line, between cycles.

the first two cycles for C_3 stimuli. The large drop from R-trial 2 to R-trial 3, without such a drop from trials 1 to 2, is strongly suggestive of the fact that Ss only recognize the common component on its second appearance and then hook the response to it on that trial. Other R-level analyses, including the correlation of P-level protocols (see p. 21), revealed no additional significant tendencies for R-level learning.

In conclusion, we have seen that a single overlap component can result in a highly significant tendency for stimuli sharing that component to be learned faster. Also, we have seen that the way in which common components are utilized changes across successive lists; however, the simple all-or-none ideas discussed in Chapters 2 and 3 prove unable to account for the pattern of results on any of the lists. Finally, a portion of the R-level analysis helped reveal the nature of the process explaining the results shown in the P-level analysis. In the hope of obtaining more errors, while still retaining the general overlap paradigm presented in this chapter, Experiment II was performed to illuminate the nature of the overlap facilitative effect discussed in Experiment I.

Experiment II

Method

The design and procedure for Experiment II was essentially the same as that for Experiment I, except for the following modifications. Twenty-one subjects were run. The data from one S was excluded, since she thought that she was supposed to write down the S-R pairs as they appeared (she was a native German and had a limited mastery of English). The first list had a 2 C_3, 2 C_0 structure (just as the first list of Experiment I); however, boys' first names were used as the components instead of animal names.

The major departure from Experiment I was to make the second two lists have 16 stimuli each. The stimuli were partitioned into 4 sets of 4 stimuli, and each set had the same structure. The structure for all sets of 4 stimuli was that 3 of the 4 stimuli shared a single common component, whereas the 4th consisted of all unique components and provided a control for the learning of the three with an overlap component. Denote by C_3^+ the three stimuli which shared a component and by C_3^- the single stimulus with all unique components. Finally, the components for Lists II and III were either animal names or names of common American cities, i.e., a random one of these two lists would have animal name components and the other one names of cities as the components.

It was hoped that, by increasing the list length from 12 to 16 and using the more difficult (established by a pilot study) city and animal names, learning would be retarded. In retrospect, this hope was only partially justified.

Results and Discussion

As was expected, the data from List I were very similar to the data from List I of the preceding experiment. This was anticipated, because both lists had a 2 C_3, 2 C_0 structure. The single important difference (which was expected) was that List I for this experiment proved easier than List I from the preceding experiment. No comprehensive analysis of the data from this list is presented here; the reader is referred to the discussion of List I for Experiment I for the major qualitative features of the data. The learning curve for this list, however, is presented in Fig. 6.8. Figure 6.8 is similar to the learning curve for List I (Experiment I) in Fig. 6.1; however, it is not quite so S-shaped. Next, we move to the analysis of lists II and III.

Unfortunately, there was still a learning-to-learn effect from List II to List III, and therefore their analysis will be carried out separately. This difference was not anticipated, since it did not occur measurably from List II to List III in the preceding experiment. Perhaps it can be attributed in part to the similarities in structure of Lists II and III. Also, the fact that the lists were longer, and hence the S got more experience from List II, and the fact that the warm-up task was easier with consequently less experience prior to List II, might have contributed to this learning-to-learn effect. Even with this necessary separation in analyses, there were 240 C_3^+ and 80 C_3^- P-level protocols for each list.

The learning curves for List I and List II are presented in Fig. 6.9, the mean total errors and mean trial number of last error in Table 6.3, the distributions of T and L in Figs. 6.10 and 6.11, respectively, and

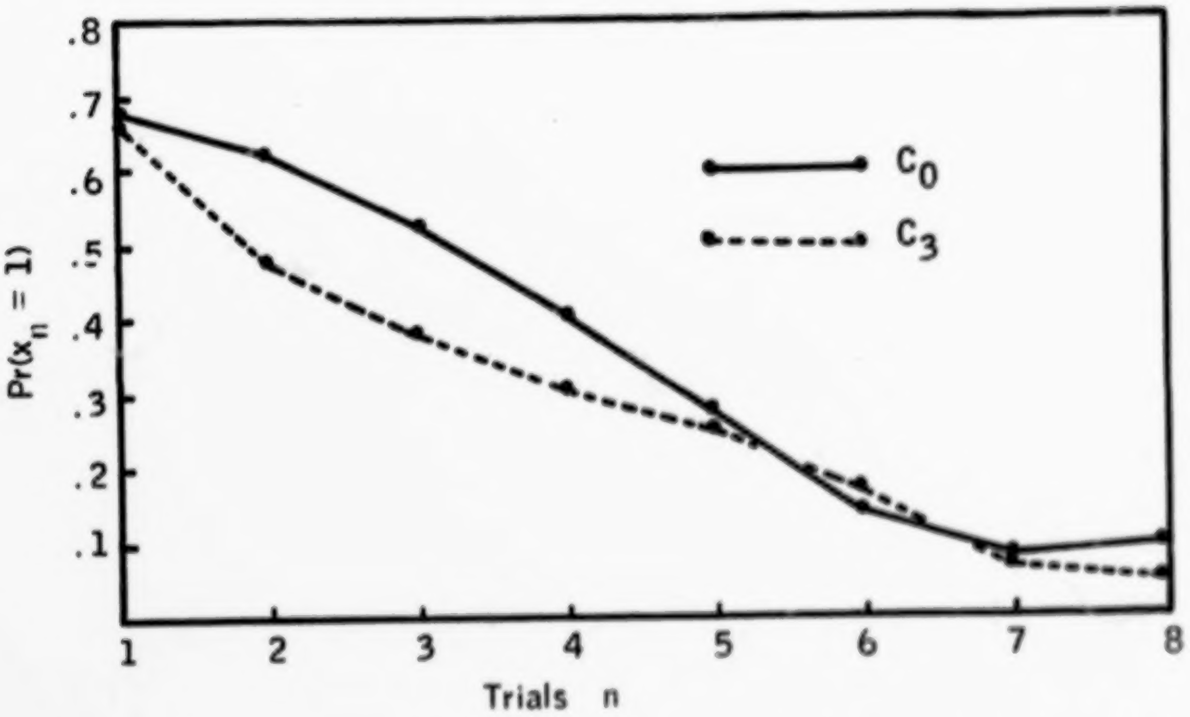


Fig. 6.8. P-level Learning Curve for C_0 and C_3 for List I (Exp. II).

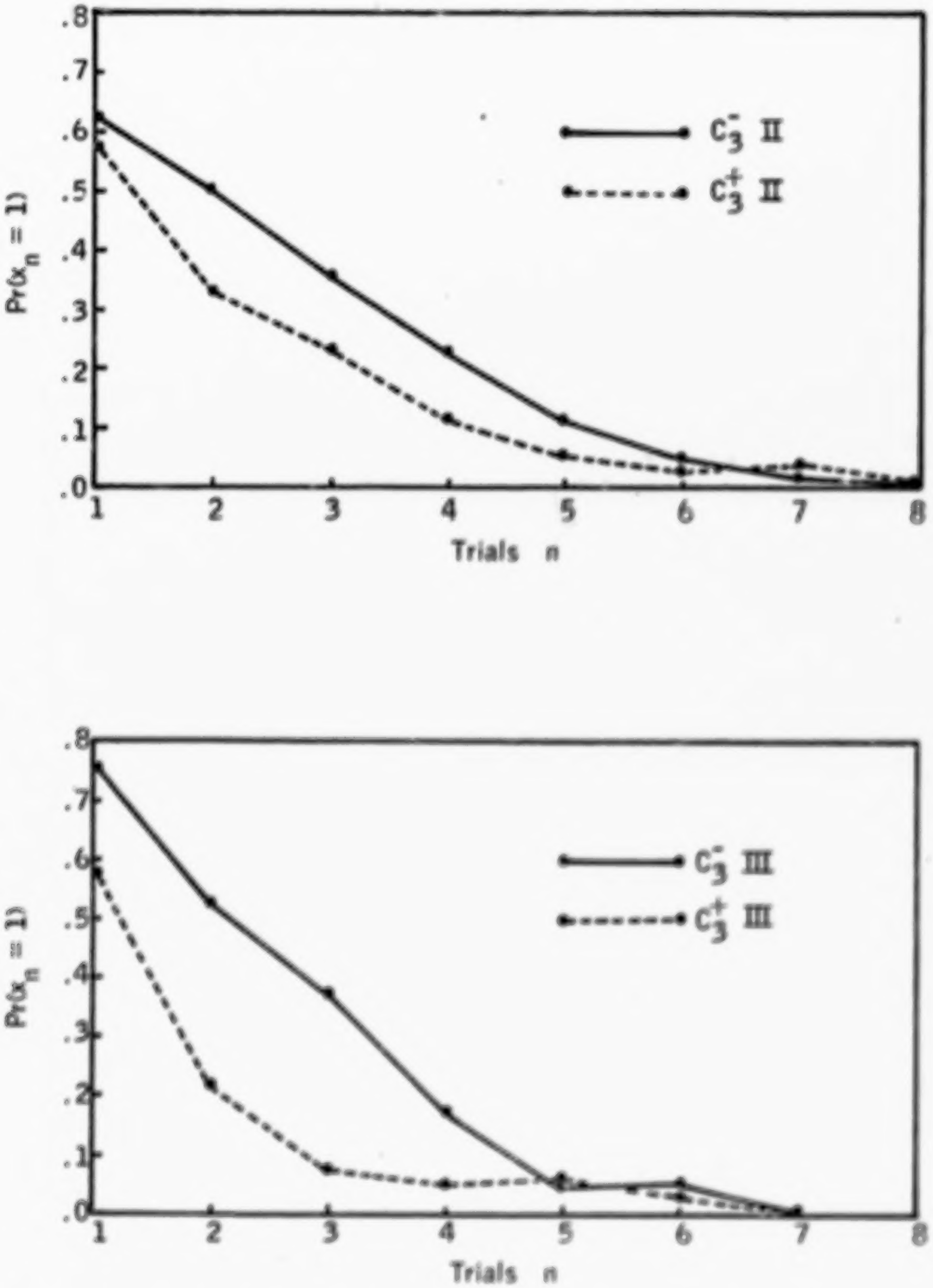


Fig. 6.9. P-level Learning Curves for C_3^- and C_3^+, Lists II, III (Exp. II).

Table 6.3. Mean Total Errors, $E(T)$, and Mean Trial Number of Last Error, $E(L)$, for C_3^+, C_3^- and Lists II, III.

List		$E(T)$	$E(L)$
II	C_3^+	1.38	1.79
II	C_3^-	1.87	2.67
III	C_3^+	0.93	1.19
III	C_3^-	.90	2.33

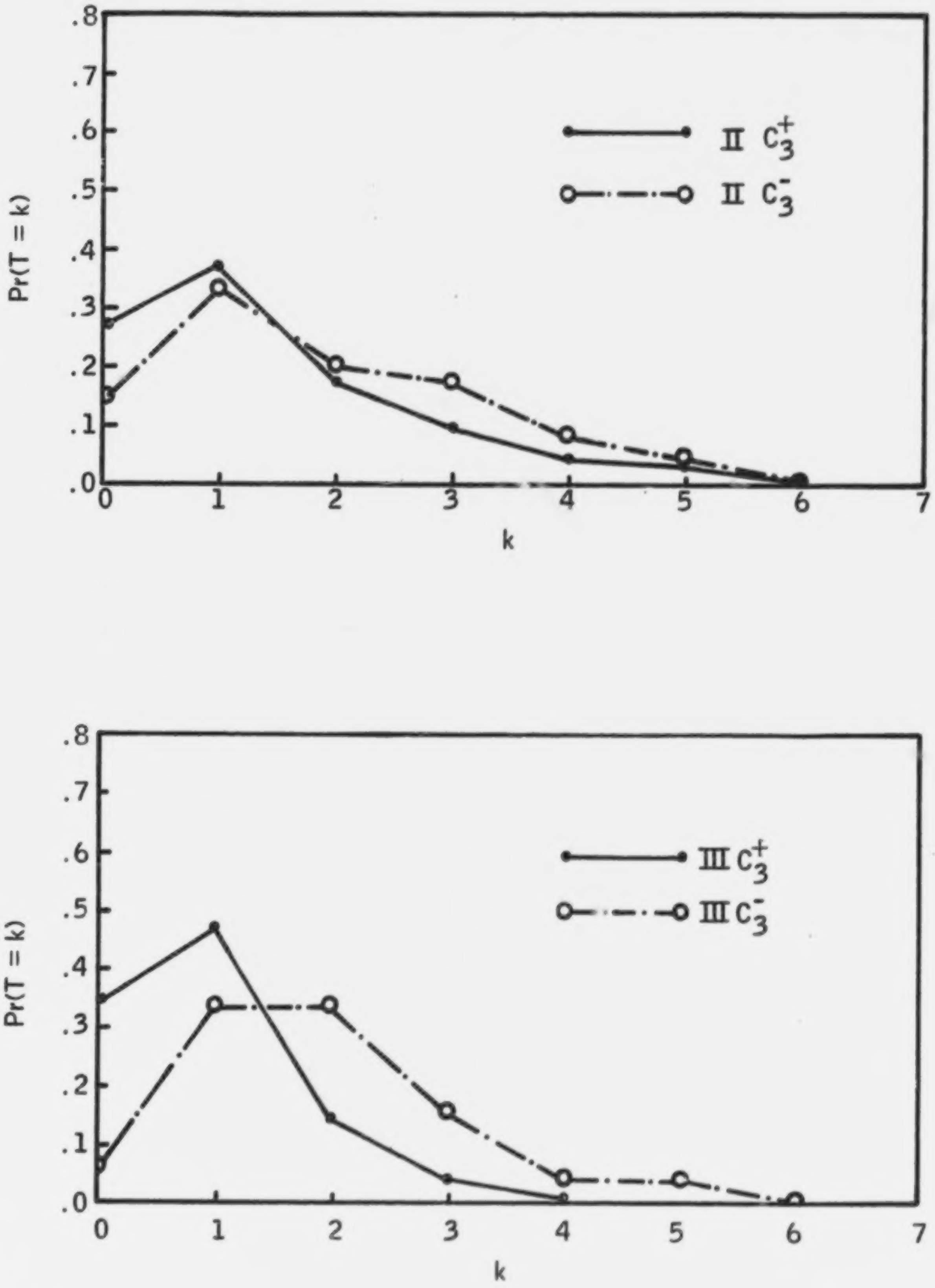


Fig. 6.10. Distributions of Total Errors, T, for C_3^+ and C_3^-, Lists II, III (Exp. II).

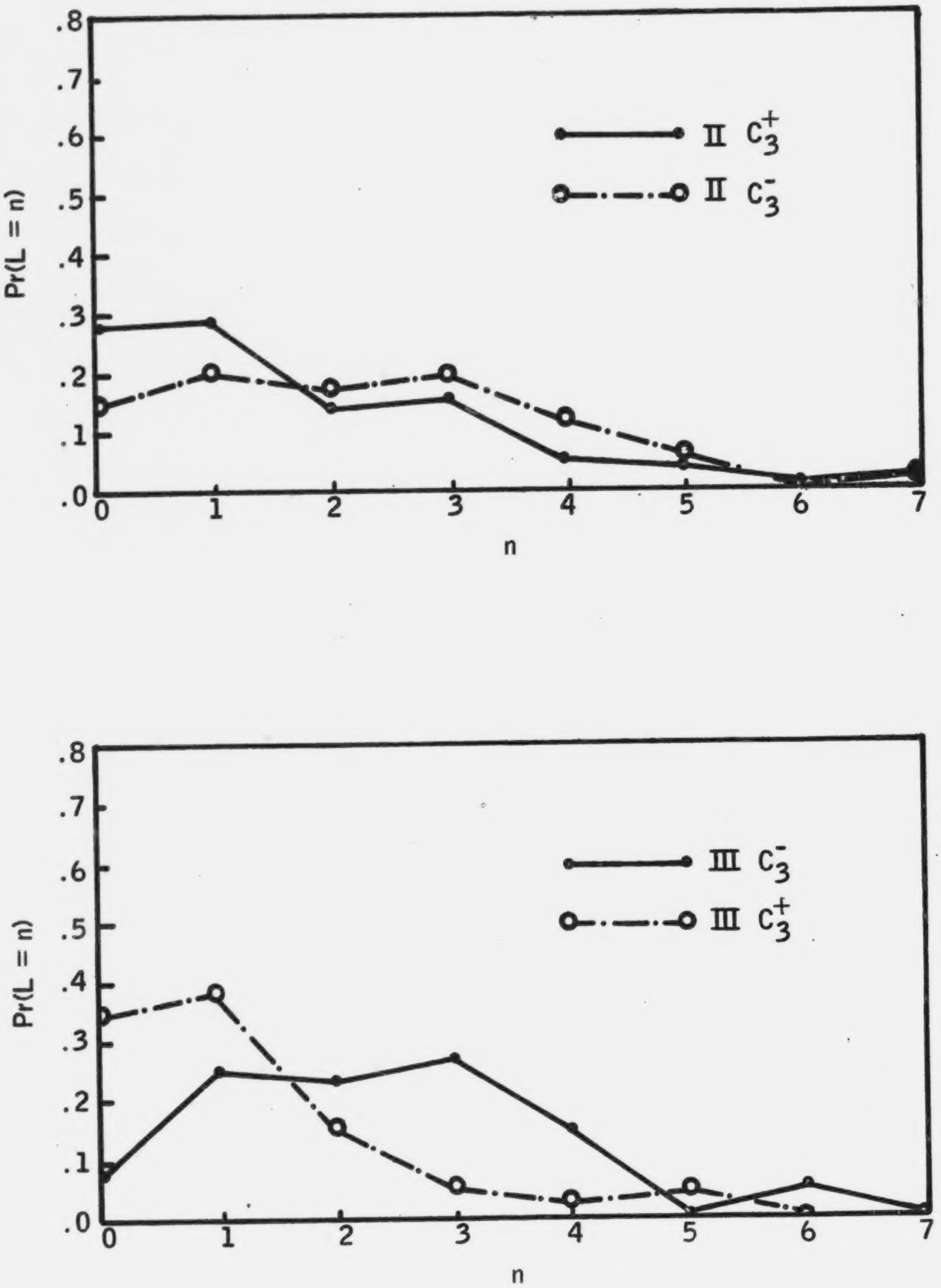


Fig. 6.11. Distributions of the Trial Number of Last Error, L, for C_3^+ and C_3^-, Lists II, III (Exp. II).

$Pr(x_n = 1|L > n)$ in Table 6.4, and $Pr(x_{n+1} = 1|x_n = 1)$ in Table 6.5. It should be emphasized that these statistics are for a P-level analysis of the data from Lists II and III.

The learning curves in Fig. 6.9 show that learning was much faster for C_3^+ stimuli than for C_3^- stimuli. Also, the curves for Lists II and III demonstrate a fairly striking learning-to-learn effect for C_3^+ stimuli, i.e., C_3^+ stimuli in List III were learned much more rapidly than they were in List II. Neither the two C_3^- nor the List III C_3^+ learning curves are exponential in shape. The C_3^- curves take about equal drops in the error probability for the first three trials and the List III C_3^+ curve drops much too rapidly from trial 1 to 2 to be approximated by an exponential function. This evidence, as well as other evidence, suggests that the data would not be fit well by a P- or R-level one-element model or the all-or-none multi-level model, since all three models imply an exponential P-level learning curve (see Chapter 2, p. 18 and Chapter 3, p. 39).

Table 6.3 presents more evidence on the learning-to-learn effect and the superiority of C_3^+ over C_3^- stimuli in learning rate. Much to the writer's chagrin, the 16-item list proved remarkably easy for the Stanford students, so it is very difficult to undertake any elaborate protocol analyses. The E(T) column in Table 6.3 shows how few errors were actually made to the stimuli. The T distributions in Fig. 6.10 reveal geometric-like distributions; however, the L distributions in Fig. 6.11 seem not to be geometric. Finally, the strongest indicator that a model with more than a single stage all-or-none feature is needed to account for these data is seen in the tendency for $Pr(x_{n+1} = 1|x_n = 1)$

Table 6.4. $\Pr(x_n = 1 | L > n)$ for C_3^+, C_3^- and Lists II, III.**

List		Trials, n 1	2	3	4
II	C_3^+	.65	.69	.59	.59*
II	C_3^-	.63	.72	.55*	--
III	C_3^+	.64	.46	.38	--
III	C_3^-	.80	.64	.67*	--

Table 6.5. $\Pr(x_{n+1} = 1 | s_n = 1)$ for C_3^+, C_3^- and Lists II, III.**

List		Trials, n 1	2	3	4
II	C_3^+	.38	.46	.31	.21
II	C_3^-	.52	.45	.40	.28*
III	C_3^+	.22	.21	.21*	--
III	C_3^-	.49	.28	.18*	--

* means two adjacent trials pooled.

** computations were made only if the number of cases was greater than 30.

to decrease with trials in Table 6.5.

Evidence for R-level learning comes from a plot of the R-level learning curve. This curve is presented in Fig. 6.12. The List II R-level learning curve shows only a slight tendency to decline within a cycle. Any significant tendency for Pr(error on R-trial N) to decrease within a cycle for C_3 stimuli can be interpreted as positive transfer to items within the class. This within-cycle decline in $Pr(x_N = 1)$ is more strikingly demonstrated in the R-level learning curve for List III. The $N = 2$ to $N = 3$ decrease in $Pr(x_N = 1)$ is especially noticeable. In both Lists II and III, the larger jumps in $Pr(x_N = 1)$ take place between cycles. These jumps are thought to reflect both R-level and P-level learning, whereas the within-cycle jumps merely reflect R-level learning. Unfortunately, there are not enough errors to warrant a further R-level analysis.

In conclusion, these experiments have shown how some of the analyses discussed in Chapter 2 can be used to infer properties of data when multi-level learning is presumed to take place. Although the effect of overlap components in learning rate is striking, the general lack of many errors by the Ss prohibited a detailed R-level analysis which might have revealed the nature of this overlap facilitation. Also, it was hoped that the all-or-none multi-level model would give a fair accounting of the data in Lists II and III. This hope failed to materialize. Since it is not the purpose of this paper to attempt *post hoc* model fits to data, no effort was made to piece together a workable model for the results. Such a model would no doubt have to involve more than one stage, because the $Pr(x_{n+1} = 1 | x_n = 1)$ and $Pr(x_n = 1 | L > n)$ curves decreased

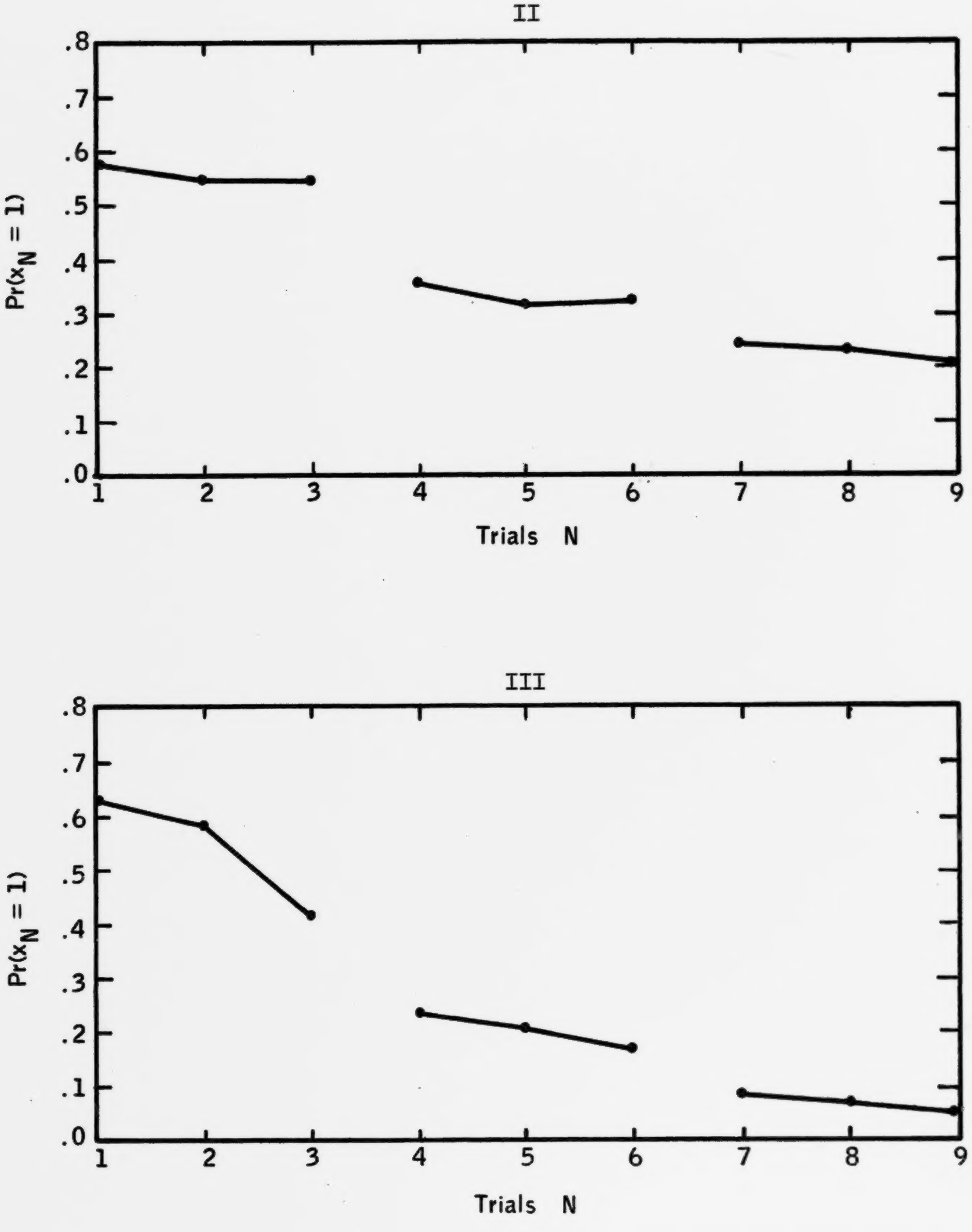


Fig. 6.12. R-level Learning Curves for the First Three Cycles on C_3^+ for Lists II, III. Solid lines are within a cycle and interruptions represent between cycles.

over trials. Also, intuition and subject interviews indicate that subjects learn to recognize overlap components before hooking responses to them, suggesting at least one additional stage. In the future, an effort will be made to find more difficult materials where overlap components do not stand out but are learned just as any other component with, perhaps, more "unconscious" positive transfer. Then a multi-level model, such as the all-or-none multi-level model, might give a better account of the data.

In addition, a program of research in which different presentation schedules for the same list are used is contemplated. It is hoped that a multi-level model written in terms of the framework in Chapter 4 can be tested by derivations from $(\mathcal{T}, \mathcal{P}, \mathcal{L})$ under various presentation schedules. The aim of this research will be to show that a valid model for paired-associate learning can be represented in a general enough way to allow for testing on various levels of data analysis as well as for different presentation schedules. It is hoped that by this line of research the ideas embodied in various extant models for P-level analyses of paired-associate learning by the anticipation procedure can be elevated to the status of a general paired-associate theory capable of making contact with data in a variety of different experimental paradigms. If this paper has contributed in any way to narrowing the apparent conceptual gap between the carefully designed simple list-learning studies of mathematical learning theorists and the more complex multi-factored processes studied by the more traditional schools of verbal learning, then it will have served its purpose.

APPENDIX I

This appendix presents selected derivations for statistics presented in Tables 2.1, 2.2, and 3.2. Not all expressions in these tables will be derived, but the hope is to convey the idea of how the multilevel derivations were made.

For Table 2.1

A) $\Pr(x_n=1)$ for the (P,R) analysis.

Since the item can appear in any of the M positions on its n^{th} cycle, the result is

$$\Pr(x_n=1) = (1-c)^{M(n-1)} \sum_{k=1}^{M} \frac{1}{M}(1-g)(1-c)^{k-1} \tag{I.1}$$

$$= \frac{(1-g)[1-(1-c)^M]}{Mc} [(1-c)^M]^{n-1} \;.$$

B) $\Pr(x_{n+1}=1|x_n=1)$ for the (P,R) analysis.

$$\Pr(x_{n+1}|x_n=1) = \frac{\Pr(x_n=1|x_{n+1}=1)\Pr(x_{n+1}=1)}{\Pr(x_n=1)} \tag{I.2}$$

$$= (1-g)(1-c)^M \;.$$

C) $\Pr(T=0)$ for the (P,R) analysis.

Let A_i be the position of the item on cycle 1, $i=1,2, \ldots , M$.

$$\Pr(T=0) = \sum_{i=1}^{M} \Pr(T=0|A_i)\Pr(A_i) \;. \tag{I.3}$$

$\Pr(A_i) = \frac{1}{M}$ and we can get $\Pr(T=0|A_i)$ in terms of i and $\Pr(T=0)$ as follows:

$$(I.4)\quad \Pr(T=0|A_i) = [1-(1-c)^{i-1}] + (1-c)^{i-1}\, g\{[1-(1-c)^{M-i+1} + (1-c)^{M-i+1}\Pr(T=0)\} .$$

Substituting (I.4) into (I.3) and summing yields

$$(I.4)\qquad \Pr(T=0) = 1 - \frac{(1-g)[1-(1-c)^M]}{Mc[1-g(1-c)^M]}$$

$$= 1 - \frac{(1-g)b}{Mc} ,$$

where b has been substituted from Table 2.1. $\Pr(T=k)$ is computed in a similar manner.

For Table 2.2

A) $\Pr(x_N = 1)$ for the (R,P) analysis.

Since knowing the R-trial allows us to find the cycle nymber, $K(N)$, we have

$$(I.6)\qquad \Pr(x_N=1) = (1-g)(1-c)^{K(N)-1} .$$

B) $\Pr(x_{N+1}=1|x_N=1)$ for the (R,P) analysis.

The only difficulty in this computation is in noting that there are two cases. In the first case the item is not the last in a cycle, and hence, the $N+1^{st}$ appearing item is some other item than the N^{th}. In the second case the item is the last in a cycle and may or may not be the $N+1^{st}$ item.

The derivations that were cumbersome or not presented involve working with the maximum of a sequence of M random variables.

For Table 3.2

Derivations for the P-level are very similar to those of Table 2.1, except more cumbersome. The only difference is that on R-trials when the item appears $(1-p-r)$ is the probability that that item remains unlearned, whereas when other items appear, the probability is $(1-r)$. To illustrate, consider the learning curve. Let $A_{k,n}$ be the event of the item appearing in position k on cycle n, $k = 1,2, \ldots, M$ and $n = 1,2, \ldots$.

$$\begin{aligned}
\Pr(x_n=1) &= \frac{1}{M}\sum_{k=1}^{M} \Pr(x_n=1 \mid A_{k,n}) \\
&= \frac{1}{M}\sum_{k=1}^{M} [(1-p-r)(1-r)^{M-1}]^{n-1} (1-r)^{k-1} (1-g) \\
&= \frac{(1-g)[1-(1-r)^M]}{Mr} [(1-p-r)(1-r)^{M-1}]^{n-1} .
\end{aligned}$$

The two derivations presented for the R-level analysis are very similar to those of Table 2.2, except that during a cycle the R-level process operates. To illustrate, consider the learning curve. Since the cycle number, $K(N)$, associated with R-trial N is easily computed, we have

$$\Pr(x_N=1) = (1-g)(1-r)^{N-1} \left(\frac{(1-r-p)}{(1-r)}\right)^{K(N)-1} .$$

REFERENCES

Atkinson, R. C. (Ed.), 1964. *Studies in mathematical psychology*. Stanford: Stanford Univer. Press.

Atkinson, R. C., G. H. Bower, and E. J. Crothers, 1965, *An introduction to mathematical learning theory*. New York: Wiley.

Atkinson, R. C. and E. J. Crothers, 1964. A comparison of paired-associate learning models having different learning and retention axioms. *J. math. Psychol.*, *1*, pp. 285-315.

Atkinson, R. C. and W. K. Estes, 1963. Stimulus sampling theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology*, Vol. II. New York: Wiley, pp. 121-268.

Atkinson, R. C. and R. M. Shiffrin, 1965. Mathematical models for memory and learning. Tech. Rep. No. 79, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Batchelder, W. H., R. A. Bjork, and J. I. Yellott, Jr., 1966, *Problem book in mathematical learning theory*. New York: Wiley.

Battig, W. F., 1966. Evidence for coding processes in "rote" paired-associate learning. *J. verbal Learn. verb. Behav.*, *5*, pp. 177-181.

Bernbach, H. A., 1966. Derivation of learning process statistics for a general Markov model. *Psychometrika*, *31*, pp. 225-234.

Bjork, R. A., 1966. Learning and short-term retention of paired associates in relation to specific sequences of interpresentation intervals. Unpublished doctoral dissertation to be published as Tech. Rep. No. 106, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Bower, G. H., 1960. Paired associates under two training conditions and different numbers of response alternatives. *Amer. Psychologist*, *15*, 451 (Abstract).

Bower, G. H., 1961. Application of a model to paired-associate learning. *Psychometrika*, *26*, pp. 255-280.

Bower, G. H. and J. Theios, 1964. A learning model for discrete performance levels. In R. C. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford: Stanford Univer. Press, pp. 1-31.

Bower, G. H. and T. R. Trabasso, 1964. Concept identification. In R. C. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford: Stanford Univer. Press, pp. 32-94.

Burke, C. J. and M. Rosenblatt, 1958. A Markovian function of a Markov chain, *Annals of Mathematical Statistics*, *29*, pp. 1112-1122.

Bush, R. R. and W. K. Estes (Eds.), 1959. *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press.

Calfee, R. C. and R. C. Atkinson, 1965. Paired-associate models and the effects of list length. *J. math. Psychol.*, *2*, pp. 255-267.

Cofer, C. N., 1961. *Verbal learning and verbal behavior*. New York: McGraw-Hill.

Cofer, C. N., 1966. Some evidence for coding processes derived from clustering in free recall. *J. verbal Learn. verb. Behav.*, *5*, pp. 188-192.

Cofer, C. N. and B. S. Musgrave, 1963. *Verbal behavior and learning*. New York: McGraw-Hill.

Cohen, B. H., 1966. Some-or-none characteristics of coding behavior. *J. verbal Learn. verb. Behav.*, *5*, pp. 182-187.

Crothers, E. J., 1963. General Markov models for learning with intertrial forgetting. Tech. Rep. No. 53, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Crothers, E. J., 1964. All-or-none learning with compound responses. In R. C. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford: Stanford Univer. Press, pp. 95-115.

Crothers, E. J., 1965a. Learning model solution to a problem in constrained optimization. *J. math. Psychol.*, *2*, pp. 19-25.

Crothers, E. J., 1965b. Optimal presentation orders for items from different categories. Tech. Rep. No. 71, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Crothers, E. J. and P. Suppes, 1967. Some experiments on learning Russian. Academic Press: New York, (in press).

Estes, W. K., 1959. Component and pattern models with Markovian interpretations. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, pp. 9-52.

Estes, W. K. and B. L. Hopkins, 1961. Acquisition and transfer in pattern-vs.-component discrimintation learning. *J. exp. Psychol.*, 61, pp. 322-328.

Estes, W. K., B. L. Hopkins, and E. J. Crothers, 1960. All-or-none and conservation effects in the learning and retention of paired-associates. *J. exp. Psychol.*, 60, pp. 329-339.

Estes, W. K. and P. Suppes, 1959. Foundations of statistical learning theory, II. The stimulus sampling model for simple learning. Tech. Rep. No. 26, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Friedman, M. P., 1966. Transfer effects and response strategies in pattern-versus-component discrimination learning. *J. exp. Psychol.*, *71*, pp. 420-428.

Friedman, M. P. and H. Gelfand, 1964. Transfer effects in discrimination learning. *J. math. Psychol.*, *1*, pp. 204-214.

Friedman, M. P., T. Trabasso, and L. Mosberg, 1966. Tests of a mixed model for paired associate learning with overlapping stimuli. Unpublished paper.

Gibson, Eleanor J., 1940. A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychol. Rev.*, *47*, pp. 196-229.

Greeno, J. G., 1966. Paired-associate learning with short-term retention: mathematical analysis and data regarding identification. Mimeographed paper.

Greeno, J. G., and T. E. Steiner, 1964. Markovian processes with identifiable states: general considerations and application to all-or-none learning. *Psychometrika*, 29, pp. 309-333.

Groen G. J., and R. C. Atkinson, in press. Models for optimizing the learning process. *Psychol. Bull.*

Izawa, Chizuko I., 1966. Reinforcement-test sequences in paired-associate learning. Doctoral dissertation published in *Psychol. Reports*, *18*, pp. 879-919.

Kemeny, J. G., and J. L. Snell, 1960. *Finite Markov chains*. Princeton: D. Van Nostrand Company.

Kemeny, J. G., Mirkil, J. L. Snell, and G. L. Thompson, 1959. Finite Mathematical Structures. Prentice-Hall. Englewood Cliffs, N. Y.

Kendler, H. H., 1966. Coding: associationistic or organizational? *J. verbal Learn. verb. Behav.*, *5*, pp. 198-200.

Luce, R. D., R. R. Bush, and E. Galanter (Eds.), 1963. *Handbook of mathematical psychology*. Vols. I and II. New York: Wiley.

Melton, A. W., 1963. Comments on Professor Peterson's Paper. In C. N. Cofer and B. S. Musgrave (Eds.), *Verbal behavior and learning*. New York: McGraw-Hill, pp. 353-370.

Millward, R., 1964. An all-or-none model for noncorrection routines with elimination of incorrect responses. *J. math. Psychol.*, *1*, pp. 392-404.

Norman, M. F., 1964. Incremental learning on random trials. *J. math. Psychol.*, *1*, pp. 336-350.

Peterson, L. R., 1963. Immediate memory: data and theory. In C. N. Cofer and B. S. Musgrave (Eds.), *Verbal behavior and learning*. New York: McGraw-Hill, pp. 336-353.

Peterson, L. R. and Margaret J. Peterson, 1962. Minimal paired-associate learning. *J. exp. Psychol.*, *63*, pp. 521-527.

Polson, M. C., F. Restle, and P. G. Polson, 1965. Association and discrimination in paired-associate learning. *J. exp. Psychol.*, *69*, pp. 47-55.

Postman, L., 1963. One-trial learning. In C. N. Cofer and B. S. Musgrave (Eds.), *Verbal behavior and learning*. New York: McGraw-Hill, pp. 295-333.

Restle, F., 1961. Statistical methods for a theory of cue learning. *Psychometrika*, *26*, pp. 291-306.

Restle, F., 1962. Conditioning and discrimination: the all-or-none process in paired-associate learning. Technical Report, Department of Psychology, Indiana Univ.

Restle, F., 1964. Sources of difficulty in learning paired associates. In R. C. Atkinson (Ed.) *Studies in mathematical psychology*. Stanford: Stanford Univer. Press, pp. 116-172.

Ruskin, A. B., in preparation. Mathematical theories of concept learning. Unpublished doctoral dissertation. Stanford University.

Shepard, R. N., 1963. Comments on Professor Underwood's paper. In C. N. Cofer and B. S. Musgrave (Eds.), *Verbal behavior and learning*. New York: McGraw-Hill, pp. 48-70.

Shepard, R. N., 1966. Learning and recall as organization and search. *J. verbal Learn. verb. Behav.*, *5*, pp. 201-204.

Shepard, R. N., C. I. Hoveland, and H. M. Jenkins, 1961. Learning and memorization of classifications. *Psychol. Monogr.*, 75, No. 13 (Whole No. 517).

Suppes, P., 1964. Problems of optimization in the learning of a simple list of items. In M. Shelly and Bryan (Eds.), *Human judgments and optimality*. New York: Wiley, 1964.

Suppes, P., and R. C. Atkinson, 1960. Markov learning models for multi-person interactions. Stanford: Stanford Univer. Press.

Suppes, P., E. Crothers, R. Weir, and E. Trager, 1962. Some quantitative studies of Russian consonant phoneme discrimination. Tech. Rep. No. 49, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Suppes, P., and Rose Ginsberg, 1963. A fundamental property of all-or-none models. *Psychol. Rev.*, *70*, pp. 139-161.

Tulving, E., 1966. Subjective organization and effects of repetition in multi-trial free-recall learning. *J. verbal Learn. verb. Behav.*, *5*, pp. 193-197.

Underwood, B. J., 1963. Stimulus selection in verbal learning. In C. N. Cofer and B. S. Musgrave (Eds.), *Verbal behavior and learning*. New York: McGraw-Hill, pp. 33-48.

Young, J. L., 1966. Effects of intervals between reinforcements and test trials in paired-associate learning. Unpublished doctoral dissertation to be published as Tech. Rep. No. 101, Institute for Mathematical Studies in the Social Sciences, Stanford University.

(Continued from inside front cover)

36 G. H. Bower. Response strengths and choice probability: A consideration of two combination rules. December 19, 1960. (In E. Nagel, P. Suppes, and A. Tarski (Eds.), Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress. Stanford Univ. Press, 1962. Pp. 400-412)

37 G. H. Bower. Application of the all-or-none conditioning model to the learning of compound responses. June 7, 1961.

38 P. Suppes and M. Schlag-Rey. Test of some learning models for double contingent reinforcement. August 15, 1961. (Psychol. Rep., 1962,

39 P. Suppes and R. Ginsberg. A fundamental property of all-or-none models, binomial distribution of responses prior to conditioning, with application to concept formation in children. September 20, 1961. (Psychol. Rev., 1963, 70. 139-161)

40 J. Theios. A three-state Markov model for learning. September 22, 1961. (Simple conditioning as two-stage all-or-none learning, Psychol. Rev., 1963, 70, 403-417)

41 G. H. Bower. General three-state Markov learning models. September 26, 1961.

42 R. C. Atkinson. A variable threshold model for signal detection. November 17, 1961.

43 R. C. Atkinson. Mathematical models in research on perception and learning. December 25, 1961. (In M. H. Marx (Ed.), Theories in Contemporary Psychology. New York: Macmillan Co., 1963. Pp. 551-564)

44 P. Suppes. Towards a behavioral foundation of mathematical proofs. January 2, 1962. (In K. Ajdukiewicz (Ed.), The Foundations of Statements and Decisions: Proceedings of the International Colloquium on Methodology of Sciences, September 18-23, 1961. Warszawa: PWN-Polish Scientific Publishers, 1965. Pp. 327-341)

45 P. Suppes and J. L. Zinnes. Basic measurement theory. March 15, 1962. (Chapter I in R. R. Bush, E. H. Galanter, and R. D. Luce (Eds.), Handbook of Mathematical Psychology, Vol. I. New York: John Wiley, 1963)

46 R. C. Atkinson, E. C. Carterette, and R. A. Kinchla. Sequential phenomena in psychophysical judgments: a theoretical analysis. April 20, 1962. (Institute of Radio Engineers Transactions on Information Theory, 1962, IT-8, S 155-162)

47 R. C. Atkinson. A variable sensitivity theory of signal detection. May 18, 1962. (Psychol. Rev., 1963, 70, 91-106)

48 R. C. Atkinson and W. K. Estes. Stimulus sampling theory. July 1, 1962. (Chapter 10 in R. R. Bush, G. H. Galanter, and R. D. Luce (Eds.), Handbook of Mathematical Psychology, Vol. II. New York: Wiley, 1963)

49 P. Suppes, E. Crothers, R. Weir, and E. Trager. Some quantitative studies of Russian consonant phenome discrimination. September 14, 1962.

50 R. C. Atkinson and R. C. Calfee. Mathematical learning theory. January 2, 1963. (In B. B. Wolman (Ed.), Scientific Psychology. New York: Basic Books, Inc., 1965. Pp. 254-275)

51 P. Suppes, E. Crothers, and R. Weir. Application of mathematical learning theory and linguistic analysis to vowel phoneme matching in Russian words. December 28, 1962.

52 R. C. Atkinson, R. Calfee, G. Sommer, W. Jeffrey and R. Shoemaker. A test of three models for stimulus compounding with children. January 29, 1963. (J. exp. Psychol., 1964, 67, 52-58)

53 E. Crothers. General Markov models for learning with inter-trial forgetting. April 8, 1963.

54 J. L. Myers and R. C. Atkinson. Choice behavior and reward structure. May 24, 1963. (Journal math. Psychol., 1964, 1, 170-203)

55 R. E. Robinson. A set-theoretical approach to empirical meaningfulness of measurement statements. June 10, 1963.

56 E. Crothers, R. Weir and P. Palmer. The role of transcription in the learning of the orthographic representations of Russian sounds. June 17, 1963.

57 P. Suppes. Problems of optimization in learning a list of simple items. July 22, 1963. (In Maynard W. Shelly, II and Glenn L. Bryan (Eds.), Human Judgments and Optimality. New York: Wiley. 1964. Pp. 116-126)

58 R. C. Atkinson and E. J. Crothers. Theoretical note: all-or-none learning and intertrial forgetting. July 24, 1963.

59 R. C. Calfee. Long-term behavior of rats under probabilistic reinforcement schedules. October 1, 1963.

60 R. C. Atkinson and E. J. Crothers. Tests of acquisition and retention, axioms for paired-associate learning. October 25, 1963. (A comparison of paired-associate learning models having different acquisition and retention axioms, J. math. Psychol., 1964, 1, 285-315)

61 W. J. McGill and J. Gibbon. The general-gamma distribution and reaction times. November 20, 1963. (J. math. Psychol., 1965, 2, 1-18)

62 M. F. Norman. Incremental learning on random trials. December 9, 1963. (J. math. Psychol., 1964, 1, 336-351)

63 P. Suppes. The development of mathematical concepts in children. February 25, 1964. (On the behavioral foundations of mathematical concepts. Monographs of the Society for Research in Child Development, 1965, 30, 60-96)

64 P. Suppes. Mathematical concept formation in children. April 10, 1964. (Amer. Psychologist, 1966, 21, 139-150)

65 R. C. Calfee, R. C. Atkinson, and T. Shelton, Jr. Mathematical models for verbal learning. August 21, 1964. (In N. Wiener and J. P. Schoda (Eds.), Cybernetics of the Nervous System: Progress in Brain Research. Amsterdam, The Netherlands: Elsevier Publishing Co., 1965. Pp. 333-349)

66 L. Keller, M. Cole, C. J. Burke, and W. K. Estes. Paired associate learning with differential rewards. August 20, 1964. (Reward and information values of trial outcomes in paired associate learning. (Psychol. Monogr., 1965, 79, 1-21)

67 M. F. Norman. A probabilistic model for free-responding. December 14, 1964.

68 W. K. Estes and H. A. Taylor. Visual detection in relation to display size and redundancy of critical elements. January 25, 1965, Revised 7-1-65. (Perception and Psychophysics, 1966, 1, 9-16)

69 P. Suppes and J. Donio. Foundations of stimulus-sampling theory for continuous-time processes. February 9, 1965.

70 R. C. Atkinson and R. A. Kinchla. A learning model for forced-choice detection experiments. February 10, 1965. (Br. J. math stat. Psychol., 1965, 18, 184-206)

71 E. J. Crothers. Presentation orders for items from different categories. March 10, 1965.

72 P. Suppes, G. Groen, and M. Schlag-Rey. Some models for response latency in paired-associates learning. May 5, 1965. (J. math. Psychol., 1966, 3, 99-128)

73 M. V. Levine. The generalization function in the probability learning experiment. June 3, 1965.

74 D. Hansen and T. S. Rodgers. An exploration of psycholinguistic units in initial reading. July 6, 1965.

75 B. C. Arnold. A correlated urn-scheme for a continuum of responses. July 20, 1965.

76 C. Izawa and W. K. Estes. Reinforcement-test sequences in paired-associate learning. August 1, 1965.

77 S. L. Blehart. Pattern discrimination learning with Rhesus monkeys. September 1, 1965.

78 J. L. Phillips and R. C. Atkinson. The effects of display size on short-term memory. August 31, 1965.

79 R. C. Atkinson and R. M. Shiffrin. Mathematical models for memory and learning. September 20, 1965.

80 P. Suppes. The psychological foundations of mathematics. October 25, 1965.

(Continued on back cover)

7LA
OCT 10 1967

(Continued from inside back cover)

81 P. Suppes. Computer-assisted instruction in the schools: potentialities, problems, prospects. October 29, 1965.
82 R. A. Kinchla, J. Townsend, J. Yellott, Jr., and R. C. Atkinson. Influence of correlated visual cues on auditory signal detection. November 2, 1965. (*Perception and Psychophysics*, 1966, *1*, 67-73)
83 P. Suppes, M. Jerman, and G. Groen. Arithmetic drills and review on a computer-based teletype. November 5, 1965.
84 P. Suppes and L. Hyman. Concept learning with non-verbal geometrical stimuli. November 15, 1965.
85 P. Holland. A variation on the minimum chi-square test. November 18, 1965.
86 P. Suppes. Accelerated program in elementary-school mathematics -- the second year. November 22, 1965.
87 P. Lorenzen and F. Binford. Logic as a dialogical game. November 29, 1965.
88 L. Keller, W. J. Thomson, J. R. Tweedy, and R. C. Atkinson. The effects of reinforcement interval on the acquisition of paired-associate responses. December 10, 1965.
89 J. I. Yellott, Jr. Some effects on noncontingent success in human probability learning. December 15, 1965.
90 P. Suppes and G. Groen. Some counting models for first-grade performance data on simple addition facts. January 14, 1966.
91 P. Suppes. Information processing and choice behavior. January 31, 1966.
92 G. Groen and R. C. Atkinson. Models for optimizing the learning process. February 11, 1966.
93 R. C. Atkinson and D. Hansen. Computer-assisted instruction in initial reading: Stanford project. March 17, 1966.
94 P. Suppes. Probabilistic inference and the concept of total evidence. March 23, 1966.
95 P. Suppes. The axiomatic method in high-school mathematics. April 12, 1966.
96 R. C. Atkinson, J. W. Brelsford, and R. M. Shiffrin. Multi-process models for memory with applications to a continuous presentation task. April 13, 1966.
97 P. Suppes and E. Crothers. Some remarks on stimulus-response theories of language learning. June 12, 1966.
98 R. Bjork. All-or-none subprocesses in the learning of complex sequences. June 22, 1966.
99 E. Gammon. The statistical determination of linguistic units. July 1, 1966.
100 P. Suppes, L. Hyman, and M. Jerman. Linear structural models for response and latency performance in arithmetic. July 29, 1966.
101 J. L. Young. Effects of intervals between reinforcements and test trials in paired-associate learning. August 1, 1966.
102 H. A. Wilson. An investigation of linguistic unit size in memory processes. August 3, 1966.
103 J. T. Townsend. Choice behavior in a cued-recognition task. August 8, 1966.
104 W. H. Batchelder. A mathematical analysis of multi-level verbal learning. August 9, 1966.